没有数据，经济舞台将苍白无力
没有数据分析，掘金又从何谈起
数据分析
从这里开始

大数据时代 小数据分析

屈泽中 编著

Broadview
www.broadview.com.cn
中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

VISIT...

LANZAROTE Caliente.COM

大数据时代
小数据分析

屈泽中 编著



電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内容简介

本书是一本大数据时代下进行小数据分析的入门级教材，通过梳理数据分析的知识点，将各类分析工具进行串联和对比，例如：在进行线性规划的时候可以选择使用Excel或LINGO或Crystal Ball。工具的应用难易结合，让读者循序渐进地学习相关工具。JMP和Mintab用来分析数据，分析的结果使用Excel、LINGO、Crystal Ball来建立数据模型，最后使用Xcelsius来动态展示数据分析的结果。书中以两个人的对话为叙述方式，场景描写多，容易进入学习状态，完全是用生动的故事和实用的案例尽可能地贴近生活和工作，让数据分析生动有趣，基本上有高中数学知识就可以理解线性规划等数据分析内容。

本书不仅介绍Excel而且介绍使用其他工具软件进行数据分析，可用来拓展互联网公司、传统企业、电商企业、管理咨询公司等各行各业从事数据分析工作的分析师和管理者对数据分析的认知，也适合初中级数据分析师或者想进入数据分析行业的有志之士参考阅读。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有，侵权必究。

图书在版编目（**CIP**）数据

大数据时代小数据分析／屈泽中编著．—北京：电子工业出版社，2015.7

ISBN 978-7-121-26469-6

Ⅰ．①大…　Ⅱ．①屈…　Ⅲ．①数据处理　Ⅳ．①TP274

中国版本图书馆CIP数据核字（2015）第142313号

责任编辑：孙学瑛

印　　刷：北京丰源印刷厂

装　　订：三河市华成印务有限公司

出版发行：电子工业出版社

北京市海淀区万寿路173信箱　邮编100036

开　　本：787×980　1/16

印　　张：22.5

字　　数：468千字

版　　次：2015年7月第1版

印　　次：2015年7月第1次印刷

定　　价：69.00元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至zlts@phei.com.cn，盗版侵权举报请发邮件至dbqq@phei.com.cn。

服务热线：（010）88258888。

序　言

笔者自2008年的一个偶然机会第1次接触“数据挖掘”（Data Mining）这个新名词以来，在数据挖掘应用相关领域度过了6年。笔者的专业是化工，整天应该与塔、釜、换热器、化学反应和物料守恒等打交道。开始接触这个专业的目的是为了利用数据分析的一些功能来优化生产运营，让企业以更高的效率、更低的成本和更好的质量运营，为此需要数据积累、数据分析和数据模型。

2008年，国内企业在数据挖掘应用中摸索起步，远不如现在大数据火热。如今大数据最火的商业应用主要集中在互联网、银行和电信等领域。基于行业应用限制，笔者无法接触到真正的大数据挖掘，但是幸运的是还是碰到了职业和兴趣的重合点。

这几年的摸索是笔者职业生涯中很重要的一段时光，因此有必要将自己一路走来的心得与体会、感悟和挫折整理出来，一则是对自己的这段职业生涯做一个交代，特别是对一路引导、鼓励和支持笔者的师友和家人；二则是合理地引导类似笔者半道出家的学习者，对数据分析有兴趣却没有深厚的统计学知识和IT功底人士，笔者相信本书的内容对于广大对数据分析应用感兴趣的初学者来说都是一种宝贵经验。在学习数据分析的道路上笔者深刻认识到一个道理，即一个成功的数据分析实践的核心因素不是数据分析技术，而是对业务理解和分析思路。这也是当初学习数据分析的初衷，初学者切不可为数据分析而分析数据。

大数据挖掘需要精通数据库、计算机编程和深厚的统计学基础，有的甚至涉及运筹学范畴，是一门复合型的应用科学。大数据的案例现在是一抓一大把，如国外典型的“啤酒与尿布”的案例，在了解数据分析之前不妨来看看几个有趣的应用案例。

（1）数据新闻让英国撤军

2010年10月23日《卫报》利用维基解密的数据做了一篇“数据新闻”，即将伊拉克战争中所有的人员伤亡情况均标注于地图之上，地图上一个红点代表一次死伤事件。用鼠标单击红点后弹出的窗口则有详细的说明，包括伤亡人数、时间和造成伤亡的具体原因。密布的红点多达39万个，显得格外触目惊心，如图0-1所示。此新闻一经刊出立即引起朝野震动，推动英国最终做出撤出驻伊拉克军队的决定。



图0-1　伊拉克战争中所有的人员伤亡情况

（2）大数据与乔布斯癌症治疗

乔布斯是世界上第1个对自身所有DNA和肿瘤DNA进行排序的人，为此他支付了高达几十万美元的费用。他得到的不是样本，而是包括整个基因的数据文档。医生按照所有基因按需下药，最终这种方

式帮助乔布斯延长了几年的生命。

（3）Google成功预测冬季流感

2009年，Google通过分析5 000万条美国人最频繁检索的词汇将其和美国疾病中心在2003—2008年间季节性流感传播时期的数据进行比较。并建立了一个特定的数学模型，最终成功预测了2009冬季流感的传播，甚至可以具体到特定的地区和州。

（4）奢侈品销售

PRADA在纽约的旗舰店中每件衣服上都有RFID码，每当一个顾客拿起一件PRADA进入 试衣间，RFID会被自动识别；同时数据会传至PRADA总部。每一件衣服在哪个城市、哪个旗舰店、什么时间被拿进试衣间和停留多长时间，数据都被存储起来加以分析。如果一件衣服销量很低，以往的做法是直接收回；如果RFID传回的数据显示这件衣服虽然销量低，但进试衣间的次数多，则说明这件衣服的下场会截然不同，或者在某个细节的微小改变就会重新制造出一件非常流行的产品。

除了国外这些经常用于商业培训课程的案例外，数据分析其实并不遥远，在国内也不乏应用。例如，共和国的开国元帅林彪就曾经依靠敏锐的数据嗅觉和军事天赋成功捣毁敌营总部。

目前国内的大部分高校还没有开设数据挖掘这门专业课程，大数据分析需要依靠庞大的数据库，即需要各专业的人士通力合作，是一个团队作业。类似笔者这种半道出家的个人学习者在不具备团队协作的条件下，可以在样本数据的分析下工夫，样本数据也可以称为“小数据”，因此本书的名称定为《大数据时代的小数据分析》。

本书主要介绍应用数据分析的一系列工具，如：Excel、LINGO、Crystal Ball、JMP、Minitab和Xcelsius等，涉及的分析有预测、风险

分析、优化求解、假设检验、相关分析、回归分析和聚类分析等。但所有这些软件都不是最新版本，如Excel使用2010版；Minitab使用的V15版。在使用软件时最重要的不是版本的最新，而是理解其功能和特点，灵活地运用。即使是Excel　2003版本，只要运用得当，同样能发挥强大的功能。很多不同功能的软件都可以完成，本书主要结合不同软件的不同特点介绍其应用。

书中涉及一些专业名词和原理，如标准差和假设检验等，本书没有给出生涩难懂的定义，而只是通俗地解释这些名词。这样做原因有二：一则作为半道出家的笔者不愿，也不会定义这些理论；二则定义这些名词或原理只会让本来就让人头疼的数据分析显得更加枯燥。如果读者需要准确理解这些专业名词，可以参考其他资料。

本书中列举的一些应用都是尽可能地贴近生活和工作，让数据分析看起来尽可能有趣一些，在排列各章节的顺序时也尽量遵循软件的功能之间的逻辑关系。

本书在每一章均会应用一些有趣的案例引出讨论的重点，其中两人按照师徒问答的形式模拟实际工作中的场景循序渐进地学习分析工具，让枯燥的数据分析显得生动一些。

本书适合的读者如下。

（1）对数据分析应用有兴趣的人士。

（2）对统计、数学和码农等深奥理论不感兴趣者。

（3）想尝试自身专业的数据分析，提高技能者。

（4）想尝试数据分析工作并寻找切入点者。

本书不适合的读者如下。

（1）喜欢拍脑袋和胸脯者。

（2）见了数据就想呕吐者。

（3）爱好SAS/R/Python等豪门软件的狂热者。

（4）统计、数学和IT专业的大牛。

（5）对数据有深刻理解的科学家。

笔者是从化工这个与数据分析无关的专业开始学习数据分析的，相信只要读者能静心地读完本书也会有所收获。但是不能指望数据分析能解决所有的问题，它不是万能的。一个成功的数据分析实践的核心因素不是数据分析技术，而是对业务的理解和分析思路。

全书的原理讲解和工具操作同步，即在操作软件的同时理解其原理；列举的案例涵盖多个行业，根据案例引出所需要讨论的知识点；然后根据知识点举一反三，串联尽可能多的数据分析入门知识；同时将介绍其适合的分析工具。

在编写本书之前笔者与人大经济论坛（http://bbs.pinggu.org/）合作开发过相关的视频培训课件，其中部分工具与本书中介绍的工具相同，有需要视频课件的读者可以试听（前3节免费）。

（1）Crystal Ball初中级课程：http://www.peixun.net/view/208.html。

（2）Crystal Ball高级课程：http://www.peixun.net/view/216.html。

（3）LINGO初级课程：http://www.peixun.net/view/251.html。

（4）Minitab初级课程：http://www.peixun.net/

view/281.html。

由于笔者的水平有限，对数据分析的理解不够透彻，加之编写时间仓促，因此书中难免会出现一些错误或不准确之处，恳请读者批评指正。

目　录

序言
第1章　知己知彼，百战不殆——风险与预测分析
　1.1　预测从世界杯开始
　1.2　手机绑定消费的秘密
　1.3　笔记本电脑出国冒险记
　1.4　慧眼识分布
　1.5　分布72变
　1.6　做最优秀的面包店长
第2章　运筹帷幄，决胜千里——效益最大化
　2.1　换个思路来数鸡
　2.2　做一个精明的农场主
　2.3　见识LINGO与Crystal Ball的威力
第3章　图个明白，精彩展现——JMP精彩图表
　3.1　图个明白——常用图形
　3.2　图个明白——树图
　3.3　图个明白——SPC图
第4章　抽丝剥茧，明察秋毫——相关分析
　4.1　假设检验——大胆假设，小心求证
　　4.1.1　小心求证——均值检验
　　4.1.2　小心求证——比例检验
　　4.1.3　小心求证——非参数检验
　4.2　相关与回归分析
　　4.2.1　相关性与第三方变量
　　4.2.2　收入与支出关系——简单线性回归
　　4.2.3　最佳口感食品配方——多元线性回归
　　4.2.4　咖啡好喝，不能多喝——非线性回归
　　4.2.5　预防心血管疾病从减肥开始——二值Logistic回归

分析
4.3　人以类聚，物以群分——聚类分析
4.3.1　美好一天从早餐开始——观测值聚类分析
4.3.2　海拔是否影响血压——变量聚类分析
4.3.3　为熊猫分类——K均值聚类分析
第5章　要里子，也要面子——数据展现的艺术
5.1　哪种水果更好卖
5.2　书店利润最大化
5.3　非诚勿扰——最佳男友模型

第1章 知己知彼，百战不殆 风险与预测分析

1.1　预测从世界杯开始

1.2　手机绑定消费的秘密

1.3　笔记本电脑出国冒险记

1.4　慧眼识分布

1.5　分布72变

1.6　做最优秀的面包店长

1.1　预测从世界杯开始

预测一般根据以往发生的事情来推断即将发生事情的风险概率，它与风险如影随形，即根据已知的风险来推测未知的风险。



【场景再现】

Mr Shu和Miss Ju是一家公司运营分析部的职业数据分析师，前者是一名入职多年精通业务的资深数据分析师；后者则是职场新人，对数据分析有强烈的爱好。但由于在学校期间未接受过专业数据分析培训，因此Mr Shu负责对Miss Ju进行入职培训。

Mr Shu和Miss Ju也都是资深球迷，自然不会错过2014年世界杯夺冠预测这样的练习机会。

Mr Shu："2014年世界杯即将开战，你觉得谁的夺冠概率比较大？"

Miss Ju："我个人超级喜欢德国队的硬朗风格，但2014年的举办地在五星上将巴西的主场，可能南美国家的机会会更大一些吧。我们球迷完全是在这里拍脑袋，没有数据支撑，完全是靠猜测啦，你觉得呢？"

Mr Shu："直观感觉也很重要吧？不仅是我们球迷在这里拍脑袋凑热闹，专门的投资公司也是不会放过世界杯这样的捞金机会，著名的投资公司高盛公司就对2014年的世界杯进行其模型分析。"

Miss Ju："真的吗？这种巨无霸分析出来的应该算有理有据

吧？分析的结果怎么样？”

Mr Shu：“高盛在推出世界杯报告预测之余，围绕历届赛事对经济和股市的影响大做文章，发表了一份长达60页的分析报告。根据该行的统计模型，4强将为巴西、德国、阿根廷和西班牙，决赛则是巴西和阿根廷之争，主办国胜。”

Miss Ju：“哇，如此TIPS，实在大路货！这个预测就好比一年一度的香港国际赛马日，评分最高的4匹马顺序归来。既无惊喜，更乏惊吓。从投机角度出发毫无刺激性可言，有什么数据支撑吗？”

Mr Shu：“高盛的经济学家通过建立数据模型分析了自1960年以来超过14 000场国际比赛，最终得出了本届世界杯的预测结果。世界杯五冠王巴西在家门口捧得第6座金杯的可能性高达48.5%，而名列第2的则是桑巴军团的宿敌阿根廷。不过，潘帕斯雄鹰夺冠的几率为14.1%，几乎只是巴西夺冠几率的1/3而已。”

Miss Ju：“那按照高盛的分析，巴西应该胜券在握了吧？不过好像也没什么惊喜，巴西本来就是夺冠大热门。”

Mr Shu：“是的，这份报告的撰写人，即高盛首席经济学家也表示：‘当然，这个结论一点也不令人惊讶。作为两支足球史上最成功的球队，巴西和阿根廷杀入世界杯决赛实至名归。但巴西在模型预测中具有如此巨大的优势还是多少让我们感到惊讶。’巴西几乎是阿根廷的3倍呢。”

Miss Ju：“其他强队的夺冠概率怎么样？”

Mr Shu：“南美双雄之外，德国队成为最大热门。其捧杯几率为11.4%，成为欧洲球队中最具冠军相的球队；西班牙（9.8%）

和荷兰（5.6%）位居这份榜单的第4名与第5名；喀麦隆、阿尔及利亚和洪都拉斯则被高盛认为夺冠完全不可能。”

Miss Ju：“作为德国的铁杆球迷，我认为德国自然是大热门。但同为欧洲强队的英国和意大利呢？要知道意大利也是四星上将啊。”

Mr Shu：“贵为欧洲传统豪门的英格兰队被高盛认为只有1.4%的几率最终捧杯，高盛分析英格兰最终将会成为小组第3无法出线。并将一场不胜灰溜溜地离开巴西，而同组中晋级下一轮的将是意大利和乌拉圭。不过英格兰小组赛同组对手意大利的夺冠几率也只不过是1.5%，与三狮军团难分伯仲。”

Miss Ju：“说的好像有点道理，这样说我也觉得巴西夺冠的可能性极大。”

Mr Shu：“当然高盛这种基于过往比赛的数据模型预测并非万无一失，4年前的南非世界杯前该公司同样预测巴西是最大热门，夺冠几率高达26.6%。但球队在1/4决赛就被淘汰，最终捧杯的是高盛当时预测的第2热门西班牙。”

Miss Ju：“还有持不同意见的分析结果吗？”

Mr Shu：“当然也有持不同意见的预测模型，如著名的风险分析公司@Risk也做出了同样的预测模型，但模拟出来的结果与高盛公司分析的概率不太一样。”

Miss Ju：“@Risk分析出来的结果怎么样？”

Mr Shu：“@Risk利用数学模型分析的结果同样是巴西夺冠，但是结果比高盛的要谨慎很多。巴西的夺冠概率第1，为17%；概

率第2则为西班牙的12%。而接下来的6强分别是瑞士（8%）、希腊（8%）、德国（7%）、哥伦比亚（7%）、阿根廷（6%）和乌拉圭（5%）。如果不考虑主场因素，德国队的夺冠概率最大为19.9%。”

Miss Ju：“Mr Shu你有自己的预测吗？”

Miss　Ju：“当然，不然怎么能自称资深球迷，看看图1-1就知道了。”



图1-1　世界杯夺冠预测

Miss　Ju：“看来每个人的分析都不太一样嘛，这主要是看中的影响因素不一样。还是章鱼哥保罗最厉害，准确率达到92%。”

Mr　Shu：“当然每场球任何一种结果的准确的概率都有1/3，这种预测个人认为都是纸老虎，不靠谱。”

Miss Ju：“好像预测玩的就是概率嘛。”

Mr Shu：“是的，就像我们在之前说的一样，预测不可能把所有的影响因素考虑周全，特别是预测世界杯这种影响因素太多的项目。”

Miss　Ju：“那除了世界杯这种不太靠谱的预测是纸老虎外，其他预测也都那么不靠谱吗？有没有生活中的案例可以让我来理解一下概率的应用。”

Mr Shu：“当然不是，我们工作生活中就有很多利用概率来决策的案例，我们来举个例子来看看概率在生活中的应用。”

1.2　手机绑定消费的秘密

【场景再现】

Mr　Shu：“Miss　Ju，你知道吗？在电信行业中也有类似章鱼哥的数据帝。他们的工作职责与你我相似，只是研究的数据不一样罢了。他们专门通过数据库收集用户的相关数据，如每月的长途电话时间和总时间等。再依次做出各种貌似的优惠套餐，以使其利益最大化。”



Miss　Ju：“哦？听起来蛮有意思的。能简单介绍一下吗？举个例子说明一下。”

Mr Shu："废话少说，我们直接上正题。Miss Ju，能告诉我你现在使用的手机是合约机还是自费机吗？"

Miss Ju："当然是合约机啦，反正是每月只需要打够198元，3年就可以免费使用iPhone。多好，难道还有比这更好的吗？"

Mr Shu："那你在购合约机之前，你的话费每月是什么水平？"

Miss Ju："100元~220元之间吧，大部分时间都是在150元左右波动，没有仔细查看和分析过话费的消费情况。"

Mr Shu："电信商赚取的就是我们不在乎小钱的心理，我们列举联通和移动两个套餐的选择方案来分析，看看二者是如何赚钱的。

假设你每月电话时长在400分钟左右，通话时间成正态分布。均值为400分钟，标准差为20。其中长途电话占比30%左右，长途电话占比呈三角形分布，最少10%，最大40%。"

Miss Ju疑惑不解地问："等等，什么是正态分布？什么是三角形分布？均值我知道，标准差又是什么？"

Mr Shu："好吧，我利用数据和图形来说明你就知道什么是正态分布了。我们模拟一下你的通话时间和长途通话时间吧，在模拟之前先了解Minitab这个软件。"

Minitab软件是现代质量管理统计的领先者，为全球六西格玛实施的共同语言，它以无可比拟的强大功能和简易的可视化操作深受广大质量学者和统计专家的青睐。Minitab统计分析软件包最初是由美国宾

夕法尼亚州立大学开发的产品，具有30多年的历史。宾夕法尼亚州立大学、威斯康星州立大学及斯坦福大学讲师Barbara Ryan博士与Thomas A. Ryan，以及Jr.博士和Brian L. Joiner博士一起于1972年创建并开发了Minitab软件。其意义是“Mini＋Tabulator＝小型＋计算机”，初衷是想让学生更加容易地学习统计学。

（1）1972年Minitab由Barbara Ryan博士等人创建并开发。

（2）1976年Minitab手册编写完成，使得计算机软件的使用和统计的入门课程实现了结合。

（3）最初的Minitab作为一个学术软件包，使统计学更富趣味性且更有意义。随着商业市场需求的增加，1983年Minitab公司成立。

（4）1995年Minitab Ltd．成立。

（5）20世纪90年代，Minitab开始专注质量市场并于2002年和2006年相继引入Quality Companion和Quality Trainer。到目前为止，Minitab已发布Windows版本的Version17，并且已在工学和社会学等领域广泛使用。

Minitab以菜单的方式构成，无须学习高难的命令，只需拥有基本的统计知识便可使用。图表支持良好，特别是与六西格玛关联为持续质量改善和概率应用提供准确和易用的工具，在GE和AlliedSignal等公司已作为基本的程序使用，也同时被许多世界一流的公司所采用，包括通用电气、福特汽车、通用汽车、3M、霍尼韦尔、LG、东芝、诺基亚和Six Sigma顾问公司。

作为统计学入门教育方面技术领先的软件包，Minitab直观且易用。它与微软的产品相兼容，基于电子表格具备全面的统计分析能力，被超过4000所高等院校所采用和超过350种教科书所引用。

本书使用Minitab V15版本讲解。

利用Minitab来模拟正态分布的操作步骤如下。

（1）产生随机正态数据

在Minitab中单击【计算】|【随机数据】|【正态】选项，如图1-2所示。



图1-2　【正态】选项

（2）参数设置

弹出显示【正态分布】对话框，设置正态分布的参数，如图1-3所示。

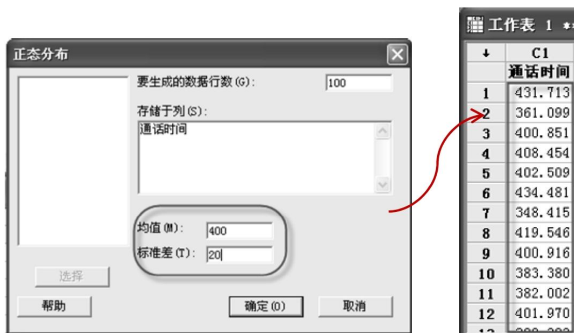


图1-3　设置正态分布参数

（3）产生随机三角形数据

单击【确定】按钮，选择【三角形分布】选项，显示【三角形分布】对话框。设置有关参数，如图1-4所示。



图1-4　设置有关参数

即每个月的长途电话时间占总通话时间的比例为0.3，最小为0.1，最大为0.4，单击【确定】按钮。

如果均值和标准差不一样，操作步骤如下。

（1）选择概率图

选择【图形】｜【概率图】选项，显示【概率分布图】对话框。选择【不同参数】选项， 如图1-5所示。

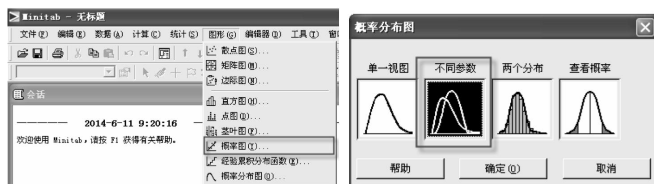


图1-5 【不同参数】选项

（2）设置不同均值

单击【确定】按钮，显示【概率分布图-不同参数】对话框。设置有关参数，如图1-6所示。单击【确定】按钮，显示【分布图】窗口，如图1-7所示。



图1-6 设置有关参数

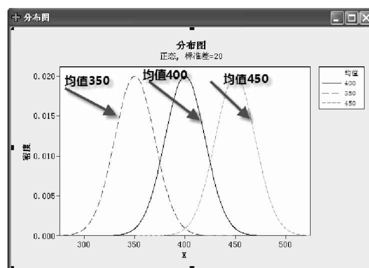


图1-7 【分布图】窗口

图中有相同的标准差，但均值不同。正态分布的概率密度函数的钟形曲线完全相同，但有一个位移。这是因为位置参数均值不一样所

致，均值大的靠右，均值小的靠左。

（3）设置不同标准差

关闭【分布图】窗口，在【概率分布图-不同参数】对话框中设置有关参数，如图1-8所示。单击【确定】按钮，显示【分布图】窗口，如图1-9所示。



图1-8　设置有关参数

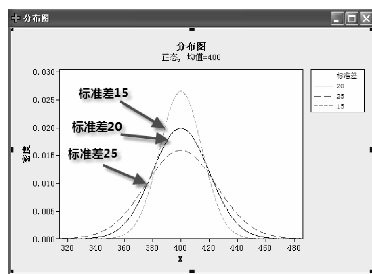


图1-9　【分布图】窗口

在相同均值和不同标准差的情况下，钟形图形的发散程度不一样。相同的均值，标准差越大的图形更为平缓且发散。正态分布的概率密度函数的总面积等于1，因此连续性图形下的面积均相等。

Miss　Ju兴奋地说："哇，正态分布好像是均值决定位置，标准差决定胖瘦嘛。"

Mr Shu答："没错，这就是概率分布的决定参数形、位、阈和尺中的两个参数。介绍其他分布后小结分布，接着看最重要的正态分布吧。"

迄今为止，在所有的概率和统计中最重要的连续分布模型就是正态分布或者称为"高斯分布"。很多其他连续性分布均可以通过正态分布变形或变换而来。连续性分布时还有许多演变分布。为了纪念高斯的伟大贡献，德国10马克的钞票上不但印上了其头像，还把正分密度曲线一起印在正面，如图1-10所示。

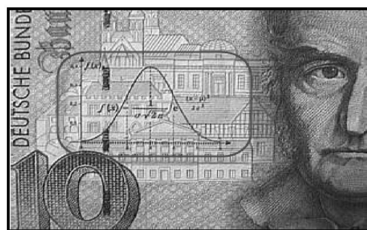


图1-10　德国10马克钞票

人们在日常生活中许多现象均服从一个正态分布，包括企业内部人员的工资水平分布、一个地区人口的身高或体重的分布，以及企业销售额的分布等。

服从正态分布的随机变量由两个参数来表征，即均值μ 和标准差σ，均值用来表征该分布的位置，即"位置参数"；标准差用来标准图形散布特征，即"尺度参数"。需要说明的是均值μ 和标准差σ 均是指总体，在实际生活中我们往往无法获得总体，因此一般从用样本均数$\bar{X}$ 和标 准差S 分别代替μ 和σ 。μ 和σ 的计算公式为：

$$\mu=\frac{1}{N}(x_1+\cdots+x_n)$$

$$\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$$

在统计学中，$\sigma=1$且$\mu=0$的正态性分布称为“标准正态分布”。

在Minitab中的操作步骤如下。

（1）选择概率图

在Minitab中单击【图形】|【概率分布图】选项，显示【概率分布图】对话框。选择【单一视图】选项，如图1-11所示。



图1-11 【单一视图】选项

（2）设置参数

单击【确定】按钮，显示【概率分布图一单一视图】对话框。如图1-12所示，设置有关参数，单击【确定】按钮，产生的标准正态分布图如图1-13所示。

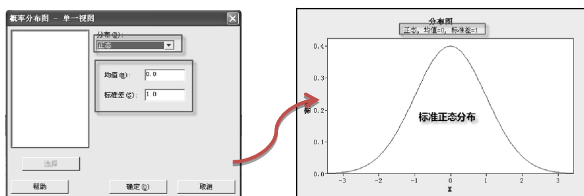


图1-12 设置有关参数

图1-13 标准正态分布图

正态分布还涉及一些非常重要的概念，如对随机数据进行标准正态化，以及根据Z值再计算概率值等，在这里暂不介绍。

Miss Ju有点得意地说："Mr Shu，你说的这些正态分布的基本概念我基本上都了解啦，这个与我们说的选择手机套餐有什么联系呢？"

Mr Shu："可不准骄傲，上面介绍的只是概率的一小部分知识。作为一名合格的数据分析师，我们要了解的是在工作中往往很多事情都是由一组数据，而不是单个数据构成。我们来看看概率分布在手机套餐选择如何应用。"

电信电话公司有两个手机套餐计划，详细情况如表1-1所示。

表**1-1**　电信电话公司的两个套餐计划

项目	套餐1	套餐2
基础话费	$39.99	$35.00
包含通话时间	400min	不限制
超时额外收费标准	$0.40	$0
长途费	$0	$0.08

从以上两个套餐可以看出两种方案各有优缺点。

套餐1的优缺点如下。

（1）优点：每月通话时间在400分钟以内，只收取基础话费$39.99，不收取长途话费。

（2）缺点：月度通话时间超过400分钟，超出的部分按每分钟加收0.04。

套餐2的优缺点如下。

（1）优点：每月不限制通话时间，只要不拨打长途电话，只收取基础话费$35.00。

（2）缺点：需要收取长途电话费，按照每分钟$0.08收取。

Mr Shu：“在这种情况下，哪个套餐更适合你呢？”

Miss Ju：“那还用说，如果长途打得多，套餐1合适；否则套餐2适合我。”

Mr Shu：“问题来了，你所指的多与少到底是多少？这就是概率的问题。还记得我前面通过Minitab中模拟通话时间的正态分布的数据吗？100行数据的均值为400，标准差为20。”

Miss Ju：“当然记的，一组100行数据就好比100个人的月度通话时间数据。但是万一有1 000个或10 000个人，岂不是要大量的计算机资源来计算？”

Mr Shu：“下面我们介绍另一款强大的数据分析工具，即Crystal Ball。”

Miss Ju：“哇哦，好好听的名字，叫水晶球。据说在欧美国家只把有魔力的东西称为‘水晶’。这款软件有什么魔力？快让我见识一下。”

Mr Shu：“不要着急嘛，让我们来了解Crystal Ball软件的一些情况。”

Crystal Ball之前是美国Decisioneering公司的产品，于1986年开发完成。Decisioneering在2007年被Hyperion公司收购，Hyperion公司之后又被Oracle收购，所以Crystal Ball目前的发行商是Oracle公司。

美国前50名最佳MBA商学院，已有40所也用Crystal Ball作为教研和商业性课题的工具。世界著名的哈佛大学商学院把Crystal Ball列为可用于计划金融的软件（Project Finance Software），因为财政计划和金融投资方面的风险分析是CB软件功能的一部分。

Crystal Ball软件与Excel

（1）Excel数据分析工具中有一项规划求解，它能解决一些简单和变量较少的线性求解。而当变量变成几十个或几百个，需要对各变量进行组合计算求解最优值时Excel显得不太便捷。例如，5个变量且每个变量有100个数据，Excel需要占用500个单元格来进行模拟计算。而借助Crystal Ball软件只需要5个单元格就可以完成，从而大大减少模型的结构。在构造更大规模且更复杂的蒙特卡洛模型时，Crystal Ball更能体现其便捷之处。Crystal Ball中没有绝对规定可以建立多少个变量，但一般来讲1 000个变量基本上能满足绝大多数模型的计算需求。

（2）Crystal Ball是Excel的一个插件，使用Crystal Ball时可以结合Excel的强大功能将数据计算推向极致。例如，在Crystal Ball中约束条件不能实现数组化处理。而此时可借助Excel中的数组处理功能，10个，甚至是百个约束条件可以变成模型中的一个约束条件。

（3）Crystal Ball与Excel都是易于上手操作的软件，仅从操作角度来讲，熟悉Excel的用户一般可以在两个小时内掌握。

Crystal Ball软件的特点

Crystal Ball帮助用户分析电子表格模型中的相关风险和不确定性，并且根据设置的假设变 量、决策变量和约束条件计算预测变量的最优解和预测分析，以及预测变量对假设变量的敏感性分析等，其主要特点如下。

（1）仿真功能强大

Crystal Ball帮助我们把日常生活生产中的事件抽象为数学模型来进行仿真计算，它可以清楚地展现因为变量变异造成模型产出的各种情况。如果没有增加仿真功能，则工作表充其量只是揭示单一结果与最一般化的情境。工作表模拟最常用的方法就是蒙特卡洛法，它可随机产生变量在不同情况下的模型结果。

（2）用户界面友好

Crystal Ball用户界面十分友好，在模型计算过程中可动态查看各类数据的变化情况，注重用户体验。

（3）停止猜测

当每个人需要规避风险时常常因为看不见的风险造成成本扩大，Crystal Ball协助用户分析所有可能，让用户握有最多信息而选择最佳方案。使用者每次使用Crystal Ball模拟时可以充分了解所处的风险。

（4）打破Excel表格限制

蒙特卡洛仿真能打破过去分析资料的能力限制，可以简单建立与分析数以万计的潜在结果时使用蒙特卡洛模拟法对某个特定状况预测所有可能的结果。运用图表对分析进行总结，并显示每一个结果的概率。蒙特卡洛是一种随机模拟方法，最早是用于研究赌术，其词汇来

源于摩洛哥赌城蒙特卡洛。

（5）了解关键控制点

透过Crystal Ball的敏感度分析可以清楚知道不同因子（假设变量）变化造成的结果，如制造业成本控制的主要因子包括几十种原料和上十种公用工程等，透过敏感性分析可以知道降低成本的关键控制点在何处。

Miss Ju：“听起来很有意思的一款软件，我们要怎样来应用它呢？”

Mr Shu：“利用它需要建立数学模型，重点要理解假设变量、预测变量和决策变量3个变量之间的关系。”

（1）预测变量：我们要做的事情，如预测世界杯冠军和哪种套餐对我更有利等。由于预测 变量时需要多个因子，因子与预测变量之间往往存在函数关系，因此预测变量有时也称为“目标函数”。

（2）决策变量：决策者可以控制的因素，其值决定预测变量的值。

（3）假设变量：决策者无法控制的变量或因素，如市场价格的走势、地震时间和每月通话时长等。

下面我们以手机套餐的案例进行数学建模，操作步骤如下。

（1）确定变量

理顺数据之间的逻辑关系，确定假设变量及预测变量。

套餐1的总话费计算逻辑：如果通话时长不超过400分钟，则收取

基本话费；否则收取额外通话时长×额外收费标准。

套餐2的总成本计算逻辑：基本话费＋总时长×长途占比×长途费。

该模型中通话时间和长途时间占总时长的比重不可控，设置为假设变量。

选择哪种套餐是我们的预定目标，但为了准确在模型中表征该变量，将套餐2的总话费与套餐1的总话费的差设置为预测变量。如其值小于0，则说明套餐2划算；反之则套餐1划算。

（2）建立Excel模型

在梳理变量后在Excel中将逻辑关系建立为数学模型。

将套餐1和套餐2的已知条件输入至Excel中。

在单元格B7中输入公式“＝IF(B9＜＝B4,B3,B3＋(B5*(B9-B4)))”，即在初始条件下套餐1的总话费。每月通话时间在400分钟以内，只收取基础话费$39.99。月度通话时间超过400分钟，超出的部分按每分钟加收$0.04，不收取长途话费。

在单元格C7中输入公式“＝C3＋(B9*B10*C6)”，即在初始条件下套餐2的总话费。每月不限制通话时间，只收取基础话费$35.00。需要收取长途电话费，按照每分钟$0.08收取。

在单元格E12中输入公式“＝C7-B7”，大于0说明套餐2要比套餐1贵，套餐1划算；反之，套餐2划算。

在Excel中建立的电话套餐模型如图1-14所示。

	A	B	C
1	电话套餐模型		
2		套餐1	套餐2
3	基础话费	$39.99	$35.00
4	包含通话时间	400	不限制
5	额外收费标准	$0.40	$0
6	长途费	$0	$0.08
7	总话费	$39.99	$44.60
8			
9	通话时间	400	
10	长途占比	30%	
11			
12	总话费之差	4.61	

图1-14　电话套餐模型

在通话时间为400分钟，长途时间占比为30%的初始条件下计算E12单元格的数值为4.61。即套餐1的总话费比套餐2少$4.61／月，套餐1优于套餐2。

如果通话时间和长途占比这两个变量有所变化，套餐2有没有可能会优于套餐1？

【问题】当“通话时长”与“长途占比”两个变量变化时对“总话费之差”变量的影响有多大？“总话费之差”又会在多大范围内波动？每次手动修改这两个变量，然后一一记录运算结果。这将会有成千上万种结果，需要很长时间才能完成，这时运用Crystal　Ball来计算和分析就会特别方便。

Excel数据模型建立完成后保存在Cell Phone.xls文件中。

（3）设置假设变量

在Excel 2010中安装的Crystal Ball软件的界面如图1-15所示。



图1-15　在Excel2010中安装的Crystal Ball的界面

打开Cell　　　Phone.xls文件，在Excel中选择B9单元格。选择【Crystal Ball】｜【Define Assumption】｜【Distribution Galley】选项，显示【Define　Assumption】窗口。选择【Normal】选项，操作过程如图1-16所示。

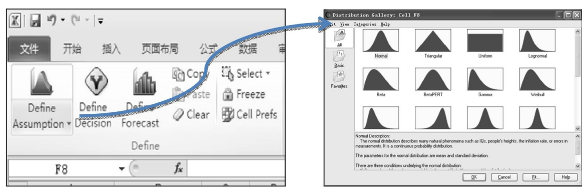


图1-16　操作过程

在弹出的窗口中设置假设变量的名称，在【Name】下拉列表框中输入“通话时间”，在【Mean】下拉列表框中输入400，在【Std.Dev】下拉列表框中输入20，表示均值为400且标准差为20的一个正态分布。

此时，B9单元格代表一组数据，即一组均值为400且标准差为20的数据形成的一个正态分布。与前面使用Minitab软件模型正态分布时一样，与Minitab不同的是此时在Excel中只需要一个单元格用于存放分布，如图1-17所示。

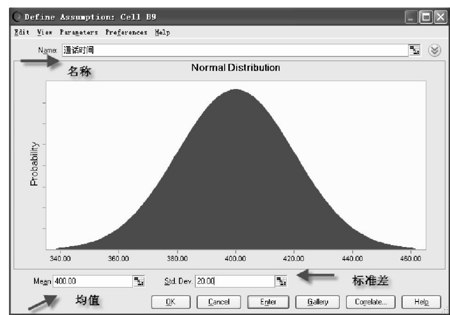


图1-17　B9单元格

选择B10单元格，在Excel中选择【Crystal Ball】|【Define Assumption】选项，选择【Distribution Galley】|【Triangular】选项。

设置假设变量的名称，在【Name】下拉列表框中输入“长途占比”，在【Maximum】下拉列表框中输入40%，在【Maximum】下拉列表框中输入10%，在【Likeliest】下拉列表框中输入30%。表示一个最大值为40%、最小值为10%且最可能值为30%的三角形分布，如图1-18所示。

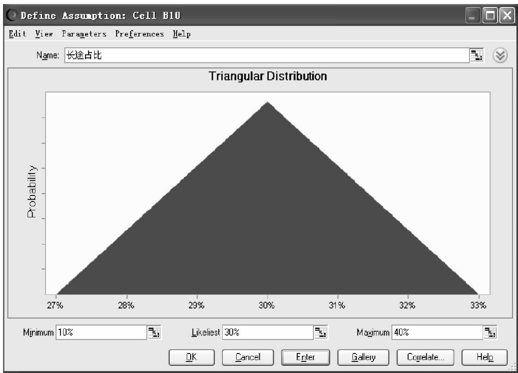


图1-18　输入值

（4）设置预测变量

选择B12单元格，在Excel菜单中选择【Crystal Ball】|【Define Forecast】选项。设置预测变量的名称，在【Name】下拉列表框中输入“总话费之差”，如图1-19所示。

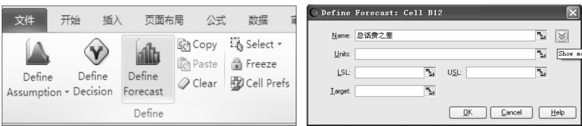


图1-19　设置预测变量的名称

到目前为止数据模型已经由普通的Excel模型转变成Crystal Ball模型，如图1-20所示。

图1-20　数学模型

在Crystal Ball软件中默认假设变量单元格为绿色底纹；选项卡和决策变量单元格为黄色底纹；预测变量单元格为淡蓝色底纹，这些单元格的底纹颜色均可以通过设置来修改。

（5）设置运行参数

在Excel菜单中选择【Crystal Ball】|【Run Preferences】选项，设置相关的运行参数。单击【Trials】标签，显示【Trials】选项卡，设置运行参数，如图1-21所示。

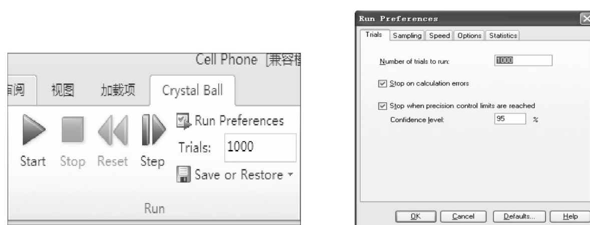


图1-21　设置运行参数

单击【OK】按钮。

（6）运行模型

在Excel菜单中选择【Crystal Ball】选项，单击【Start】按钮计算模型，如图1-22所示。

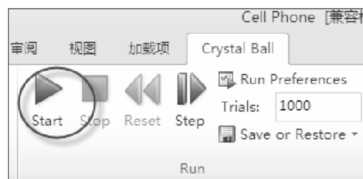


图1-22　【Start】按钮

运行后系统会自动将运算结果以直方图的形式呈现，如图1-23所示。

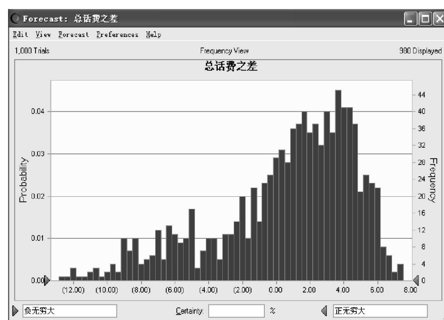


图1-23　运行结果

（7）解读结果

从结果可以看出在给定的预测条件下计算结果大部分集中在正数部分，即套餐1比套餐2优惠的概率要大，每月优惠的范围大概在$2～$5之间。

·查看累计概率

在【总话费之差】窗口中单击【View】|【Cumulative Frequency】或【Reverse Cumulative Frequency】选项，可以分别查看累计概率及逆累计概率，如图1-24所示。

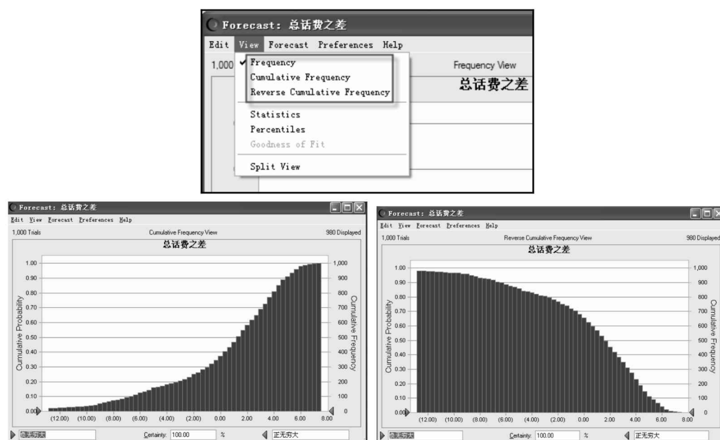


图1-24　累计概率及逆累计概率

·查看套餐1划算的定量概率

如果想查看预测变量小于0的概率，则在【总话费之差】窗口的左下单元格中输入0或直接拉动小三角形，如图1-25所示。

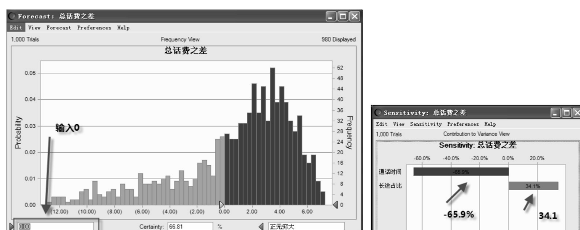


图1-25　查看套餐1划算的定量概率

可以看出预测变量“总话费之差”在0～正无穷之间的概率（两个三角形之间）为66.81%，即在给定的假设变量的条件下套餐1划算的可能性为66.81%。

·查看敏感度

在【Crystal Ball】中选择【View Charts】|【Sensitivity】选项即可以浏览敏感性分析图。

从敏感性分析来看，假设变量“通话时间”对预测变量的影响为-65.9%且“长途占比”对预测变量的影响为34.1%。二者的绝对值之和为1，显然通话时间比长途占比对预测变量影响大一些。

（8）修订参数

Miss Ju：“67%的概率好像也不能说明套餐1就一定合适我啊？说不定我就是那33%中的一员呢？”

Mr Shu：“也许你觉得67%可能还不能说明问题，那么我们将通话的分布修改为均值为430且标准差为15的分布看看结果会怎么样？”

为重新设置假设变量概率分布的参数，选择B9单元格，在Excel菜单中选择【Crystal Ball】|【Define Assumption】选项。选择【Distribution Galley】|【Normal】选项，修改参数，如图1-26所示。

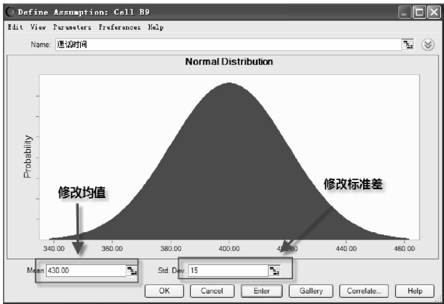


图1-26　修改参数

在Excel菜单中选择【Crystal Ball】选项，单击【Start】按钮计算模型，结果如图1-27所示。

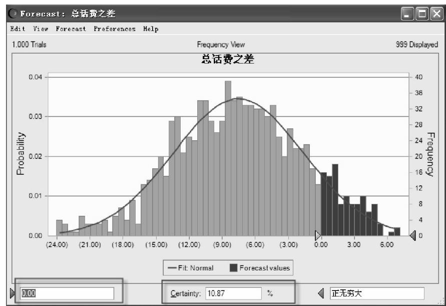


图1-27　计算结果

套餐1比套餐2划算的可能性仅为10.8%，此时我们毫不犹豫选择套餐2。即在每月通话总时间较长的情况下，套餐2的优势更加明显，这在我们前面定性分析套餐优劣势时也提到过。

对比这两种情形我们可以看出电话通话时间的分布不同可以导致完全不同的结果，套餐1划算的概率从67%骤降至11%。现在仅仅是修正了一个假设变量的分布参数，如修正“长途占 比”的分布参数计算结果可能又会不同，可见数据的准确收集和整理是多么重要。

通过对比分析我们大致能了解“通话时长”和“长途占比”这两个变量对预测变量（套餐1总费用减套餐2总费用）的影响，“通话时长”这个变量应是主导地位且是负影响。即通话时间越长，套餐2的优势越明显。

与前面查看敏感性分析一样，参数变化后同样可以非常方便查看敏感性分析图，操作步骤如下。

（1）敏感性分析图路径

在Excel菜单中选择【Crystal Ball】选项，单击【View Charts】|

【Sensitivity Charts】选项，如图1-28所示。

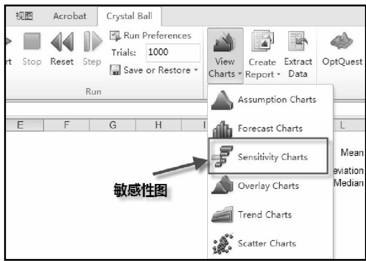


图1-28　【Sensitivity Charts】选项

（2）查看图形

选择预测变量【总话费之差】复选框，单击【Open】按钮，敏感性分析结果如图1-29所示。

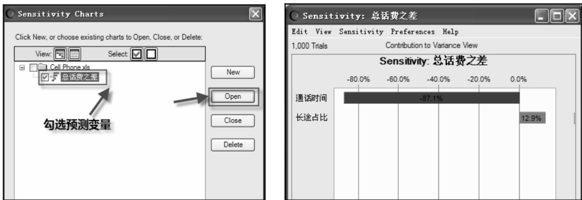


图1-29　敏感性分析结果

从敏感性结果也能看出通话时间影响预测变量的系数为-87.1%，而长途占比仅为12.9%。

Miss　Ju："哇，从这个分析来看就很明显了嘛。如果每个月的通话时间在430分钟以上，长途占比在0.1～0.4之间就应该选套餐2。可以想办法查出一个人通话数据后分析，如果是很多人的通话数据，则从哪里获取呢？"

Mr Shu："这就是电信商的优势所在，它们掌握大量的数据，

可以对用户进行聚类和关联。如把有相似消费水平的人聚成一类，然后有针对性地推出各类套餐。例如，假设大部分人的通话时间均处在均值为400和长途占比为0.3的正态分布中，此时将套餐1与iphone等手机绑定消费，你还会选择套餐2吗？而在大量用户在选择套餐1时，用户群长年累月上交的最低消费的浪费部分就足够通信公司赚得盆满钵满。”

Miss Ju恍然大悟地说：“原来秘密在这里，看来下次选择时还是要好好地分析看看。我们上面的数据模型预测变量为‘总话费之差’，两个假设变量为‘通话时长’和‘长途占比’。那么预测世界杯冠军这样比较浩大的工程，它的假设变量应该比较多吧。”

Mr Shu：“没错，如我们上面说的著名的投资公司高盛，它对世界杯的预测就是以6个因子（随机因素、榜单排名、赢球数、历届世界杯表现、主场国效应和主场大洲效应）来进行的。概率分布不再是常用的正态分布，而是离散型的泊松分布；同样也运用蒙特卡洛模拟对不同的情景概率进行测算，如图1-30所示。

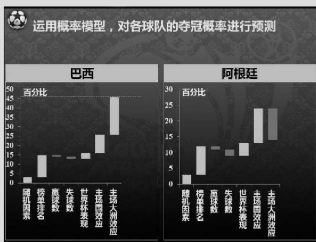


图1-30　高盛公司预测世界杯示例

Miss Ju：“看来概率真的是运用比较多！Crystal Ball这款软件好像真的蛮有意思，它还有哪些应用？”

Mr Shu：“它的应用主要是风险分析和优化计算，我们刚刚说的预测其实是风险分析的一种。未来有风险所以需要预测，我们再来看它在风险分析方面的一些应用吧。”

1.3 笔记本电脑出国冒险记

Miss Ju：“冒险记？这个标题还蛮有意思的，你不会就是标题党吧？”

Mr Shu：“这一小节我们主要讨论Crystal Ball在风险分析上的一些应用，分析的是笔记本电脑在出口之后被退运的风险分析。”

Miss Ju：“质量不合格，当然就会退运回来。”

Mr Shu：“不仅是质量不合格，还有其他很多因素都会造成产品被退运，先看看案例背景吧。”

近期国内某市出口国外的笔记本电脑退运比较频繁，海关工作人员感到比较困惑。为了搞清楚导致出口的笔记本电脑被退运的原因，他们收集到近期部分公司的笔记本电脑出口及退运的数据，如图1-31所示。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	国　家	品质原因			贸易原因			售后维修			无理由退货		
2	国　家	批次	数量（台）	额（美元）	批次	数量（台）	额（美元）	批次	数量（台）	额（美元）	批次	数量（台）	额（美元）
3	总计	845	91763	24584828	275	44913	16918691	358	3846	877165.6	10	140	175306.7
4	阿联酋 汇总	5	509	113343.5	0	0	0	0	0	0	0	0	0
6	爱尔兰 汇总	0	0	0	0	0	0	0	0	0	0	0	0
8	澳大利亚 汇总	4	3548	941865.4	15	1609	505267.9	0	0	0	0	0	0
14	巴西 汇总	10	3117	1220956	0	0	0	0	0	0	0	0	0
17	白俄罗斯 汇总	0	0	0	1	6	4654.1	0	0	0	0	0	0
19	波兰 汇总	0	0	0	8	5246	3907261	0	0	0	0	0	0
23	德国 汇总	28	1351	621150.5	1	1	574	36	311	71165	0	0	0
28	俄罗斯联邦 汇总	1	1	1043.49	0	0	0	0	0	0	0	0	0
30	法国 汇总	1	4	2289.16	3	5	1205.2	0	0	0	0	0	0
33	菲律宾 汇总	1	1	510.42	12	261	70826.33	0	0	0	0	0	0
38	斐济 汇总	0	0	0	1	1	208.64	0	0	0	0	0	0
40	芬兰 汇总	1	1	1059.22	1	1	211.52	0	0	0	0	0	0
43	韩国 汇总	4	34	21159.34	14	82	48346.93	88	782	201130	0	0	0
50	荷兰 汇总	8	3404	1339479	19	2429	1106576	0	0	0	0	0	0
55	加拿大 汇总	3	4	1327.36	0	0	0	0	0	0	0	0	0
59	捷克 汇总	0	0	0	2	2	594.54	0	0	0	0	0	0
61	捷克共和国 汇总	1	1	429.07	0	0	0	0	0	0	0	0	0

图1-31　笔记本电脑退运数据

在查看这些国家退运的数据后海关人员发现被退运的地区大部分

集中在欧洲市场，商品被退运时境外的海关部门会以“品质问题”、“贸易原因”、“售后服务”和“无理由退货”等理由来处理查看数据表。因为“无理由退货”的笔记本几乎没有，退运主要集中在“品质问题”和“贸易原因”。

境内出口的供应商都经过层层筛选，以及严格的质量认证，出口笔记本供应商在国内完全免检。境内海关日常也会对这些企业进行抽检且产品质量完全合格，因此境内海关人员也有理由相信退运是出口国家可能出于贸易壁垒的结果，但他们目前没有足够的证据或数据支撑来说 明这一点。退运一次会将一批货物全部退回来，这给企业带来很大的损失，增加了产品的物流成本。因此他们想通过蒙特卡洛模拟抽样方法及运用计算机软件来模拟企业产品被误退运而造成损失的可能性的大小，通过模拟来收集数据为商务谈判做好准备。

Miss Ju：“这个案例是不是就是说产品本身质量不错，但是现在被退运的概率比较高。而我们通过蒙特卡洛模拟来证明在这种情况下的风险并没有实际的那么高，从而来证明境外海关有贸易保护的嫌疑？”

Mr Shu：“没错，就是要对出口的这个过程进行蒙特卡洛模拟来进行风险分析。”

这里有如下假设。

（1）出口的笔记本电脑在境外过关时均按照严格的流程进行，产品有可能抽检或不抽检。

（2）每家公司每批出口的商品数量不一样，产品的合格率也是一项概率分布。到境外后每个国家对于是否抽样也是一个概率分布，每次抽样的数量也是一个概率分布，即模型中包含“产品合格率”、“是

否抽样”和“抽样数量”等3个假设变量。

基于以上假设，我们可以通过蒙特卡洛模拟定量地计算出到每个国家的产品被退运回来的风险值。

Miss Ju：“境外的海关怎么样来确定批次产品应被退货呢？”

Mr Shu：“境外的海关检验非常严格，他们会认为抽样一个样品不合格就整批退运。”

Miss　Ju：“需要产品本身质量合格率非常高，被退运的概率才会低吧？”

Mr Shu：“没错，但是我们也要注意一个问题。即质量好的也有可能被检验成质量不好的产品；质量差的也有可能会被检验成质量好的产品，检验本身也存在风险。”

Miss Ju：“也是哦，好像很复杂的样子。”

Mr Shu：“也没有，我们只要把逻辑关系梳理清晰，就可以很方便地建模。其实我们刚刚说的质量好的被判定为质量不好的可能性在统计学上称为‘一类风险’，通常可以理解为是出口企业需要承担的风险；质量差的也有可能会被检验成质量好的可能性在统计学上称为‘二类风险’，通常可以理解为客户需要承担的风险。下面我们一起来理顺思路，首先搭建这个Excel模型。”

（1）确定货物中不合格产品数

打开“手提电脑.xls”文件，以“仁宝集团”为例说明。假设仁宝集团出口到土耳其的每批次的数量为300台（先进行单值设置）且产品合格率为99%（先设置单个初值），那么不合格笔　记本的电脑台数为台数乘不合格率。

由于电脑都是整数，因此在F3单元格中输入公式“＝CEILING(D3*(1-E3),1)”对不合格的电脑进行取整。在Excel中的设置如图1-32所示。

A	B	C	D	E	F
公司	出口国家	批次	数量	产品合格率	不合格产品数量
仁宝集团	土耳其	1	300	99.00%	3

图1-32　在Excel中的变量设置

（2）设置是否检验

在G3单元格内设置是否对该批产品进行抽检，如果抽检，则设置为1；否则设置为0。

如果不抽检，则该批次的产品被退运的概率理论上为0；否则需要检验。在G3单元格中设置初值为1，设置为检验。

（3）计算不合格品被抽中的概率

注意这里设置的是不合格品被抽中的概率，不合格品被抽中不一定完全认定为不合格品。

在J3单元格中设置每次抽样的电脑台数的初值为5台，每批的产品为300台。产品的合格率为99%，则不合格的数量为3台，合格的数量为297台，全部抽中为合格产品的概率为300台中抽5台的排列组合与297台中抽5台的排列组合的比例。设置公式为COMBIN(D3-F3,J3)/COMBIN(D3,J3)，抽中一台电脑不合格的概率与抽中一台合格电脑的概率为1。因此K3单元格中设置的公式为：＝1-COMBIN(D3-F3,J3)/COMBIN(D3,J3)，表示抽中一台不合格电脑的概率，如图1-33所示。

公司	出口国家	批次	数量	产品合格率	不合格产品数量	是否抽检	次品被检验成次品概率	正品被检验成正品概率	每次抽样数量	抽样中至少有一件不合格品被抽中的概率
仁宝集团	土耳其	1	300	99.00%	3	1	0.9	0.99	5	4.9%

图1-33　不合格产品抽中概率示例

从Excel表格中可以看出在初始条件下300台笔记本中有3台不合格，每次抽5台产品检验，其中有一台为不合格品被抽中的概率为4.9%。

（4）设置检验的准确性

前面我们提到过无论质量好坏，都有可能被检验成正品或次品。我们现在关心的是无论是 正品还是次品被检验成次品的概率，因此我们假设次品被检验成次品的概率为0.9；正品检验成正品的概率为0.99。即次品被检验成正品的概率为0.1，正品被检验成次品的概率为0.01，将这两个概率值设置为常量分别输入至H3和I3单元格中。

（5）计算被退运的概率

计算的原则为抽中的产品不管是次品还是正品被检验成次品的概率为“抽中正品概率×0.01＋抽中次品概率×0.9”，所以在M3单元格中设置公式L3×(1-I3)＋(1-L3)×H3。

之前我们有假设如果不检验，则理论上不会被退运。即被退运的概率值为0，因此将M3单元格内的公式修正为“＝IF(G3＝1,L3×(1-I3)＋(1-L3)×H3,0)”。

这样可以计算出产品被退运的概率值，如图1-34所示。

公司	出口国家	批次	数量	产品合格率	不合格产品数量	是否抽检	次品被检验成次品概率	正品被检验成正品概率	每次抽样数量	抽样中至少有一件不合格品被抽中的概率	全部合格产品被抽中的概率	产品被退运概率
仁宝集团	土耳其	1	300	99.00%	3	1	0.9	0.99	5	4.9%	95.1%	5.4%

图1-34　产品被退运概率值

在产品合格率为99%且出口300台笔记本电脑，每次抽样5台笔记本，只要有一台不合格就退运的设置条件下产品被退运的概率值为5.4%。

（6）对其他出口国家进行相同设置

按照以上的逻辑对出口到每个国家进行设置，按照相同逻辑设置的被退运的概率在Excel表中一致，如图1-35所示。

公司	出口国家	批次	数量	产品合格率	不合格产品数量	是否抽检	次品被检验成次品概率	正品被检验成正品概率	每次抽样数量	抽样中至少有一件不合格品被抽中的概率	全部合格产品被抽中的概率	产品被退运概率
仁宝集团	土耳其	1	300	99.00%	3	1	0.9	0.99	5	4.9%	95.1%	5.4%
仁宝集团	日本	1	300	99.00%	3	1	0.9	0.99	5	4.9%	95.1%	5.4%
仁宝集团	印度	1	300	99.00%	3	1	0.9	0.99	5	4.9%	95.1%	5.4%
仁宝集团	阿聯酋	1	300	99.00%	3	1	0.9	0.99	5	4.9%	95.1%	5.4%
仁宝集团	俄羅斯	1	300	99.00%	3	1	0.9	0.99	5	4.9%	95.1%	5.4%
仁宝集团	南非	1	300	99.00%	3	1	0.9	0.99	5	4.9%	95.1%	5.4%
仁宝集团	美國	1	300	99.00%	3	1	0.9	0.99	5	4.9%	95.1%	5.4%
仁宝集团	英國	1	300	99.00%	3	1	0.9	0.99	5	4.9%	95.1%	5.4%
仁宝集团	香港	1	300	99.00%	3	1	0.9	0.99	5	4.9%	95.1%	5.4%
仁宝集团	挪威	1	300	99.00%	3	1	0.9	0.99	5	4.9%	95.1%	5.4%
仁宝集团	泰國	1	300	99.00%	3	1	0.9	0.99	5	4.9%	95.1%	5.4%
仁宝集团	紐西蘭	1	300	99.00%	3	1	0.9	0.99	5	4.9%	95.1%	5.4%
仁宝集团	馬來西亞	1	300	99.00%	3	1	0.9	0.99	5	4.9%	95.1%	5.4%
仁宝集团	荷蘭	1	300	99.00%	3	1	0.9	0.99	5	4.9%	95.1%	5.4%
仁宝集团	智利	1	300	99.00%	3	1	0.9	0.99	5	4.9%	95.1%	5.4%
仁宝集团	新加波	1	300	99.00%	3	1	0.9	0.99	5	4.9%	95.1%	5.4%
仁宝集团	瑞士	1	300	99.00%	3	1	0.9	0.99	5	4.9%	95.1%	5.4%
仁宝集团	臺灣	1	300	99.00%	3	1	0.9	0.99	5	4.9%	95.1%	5.4%

图1-35　产品被退运概率示例

可以看出影响产品被退运概率的因素如下。

（1）出口的批次数量，数量越大，不合格的产品也多且被抽检到的概率也就越大。

（2）产品的合格率越低，被退运的概率越大。

（3）是否检验，检验的次数越多，被退运的概率越大。

（4）每次抽样的个数越多，被退运的概率越大。

（5）产品检验的准确性越低，被退运的概率越大。

为了更好地理解模型，我们暂时把产品检验准确性设置为固定常数，因此我们还需要设置的假设变量有“批次数量”、“产品合格率”、“是否检验”、“抽样个数”。

整理完Excel逻辑关系后需在Crystal Ball中设置。首先设置假设变量，操作步骤如下。

（1）设置每批次的“出口数量”变量

选择D3单元格，在Excel菜单中选择【Crystal　Ball】选项。单击【Define　Assumption】|【Distribution　Galley】选项。选择【Normal】选项，设置有关参数，如图1-36所示。

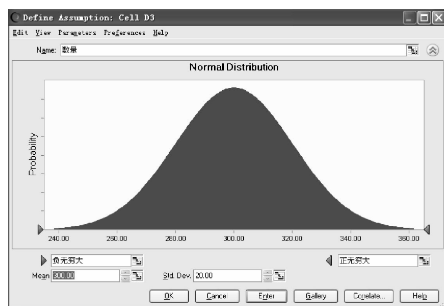


图1-36　设置有关参数

（2）设置“产品合格率”变量

选择E3单元格，在Excel菜单中选择【Crystal　Ball】|【Define　Assumption】选项。选择【Distribution　Galley】选项，选择【Normal】选项。设置有关参数，如图1-37所示。

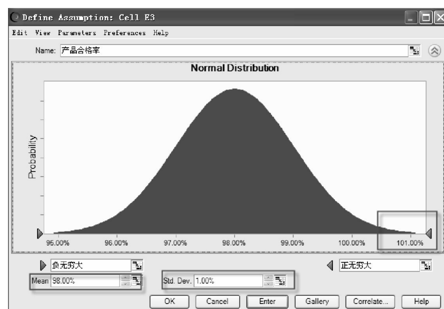


图1-37　设置有关参数

正态分布图形是一个95%～101%的概率分布，但产品合格率不可能大于100%。设置合格率最大值为99.9%，如图1-38所示。

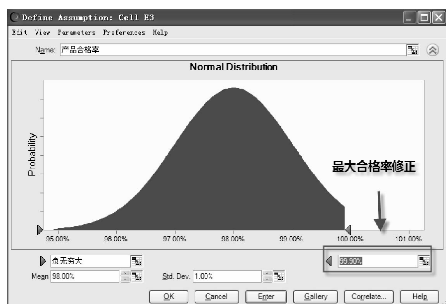


图1-38　产品合格率修正示例

（3）设置批次“是否检验”变量

检验与否只有两种可能，即检验和不检验。该变量不再是正态分布，而是0和1分布。选择G3单元格，在Excel菜单中选择【Crystal Ball】|【Define Assumption】选项。选择【Distribution Galley】选项，选择【Yes-No】选项。设置有关参数，如图1-39所示。

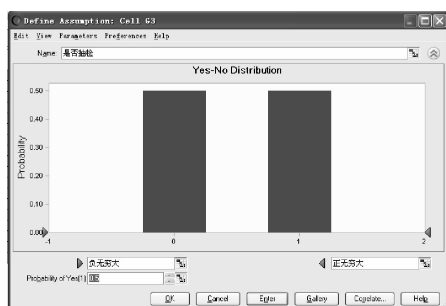


图1-39　设置有关参数

表示检验与不检验的概率相等，均为50%。

（4）设置每次“抽样数量”变量

设置每次抽样的数量为1个～5个，其抽样概率相同，该变量设置为离散型的均匀分布。选择G3单元格，在Excel菜单中选择【Crystal Ball】|【Define Assumption】选项。选择【Distribution Galley】选

项，选择【Discrete Uniform】选项。设置有关参数，如图1-40所示。

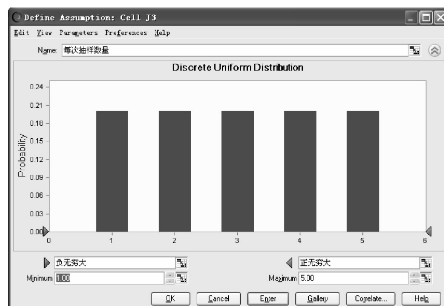


图1-40　设置有关参数

（5）设置预测变量（目标函数）“退运概率”

选择M3单元格，在Excel菜单中选择【Crystal Ball】|【Define Forecast】选项。设置预测变量名称为“产品被退运概率”，输入名称，如图1-41所示。

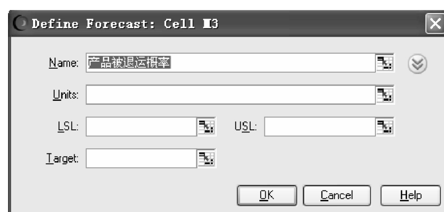


图1-41　输入名称

出口土耳其的退运模型建立完成，需要设置其他出口国家。如果重复以前操作，则比较烦琐。在Crystal Ball中可以快速复制变量设置。

（6）快速设置其他出口地的参数

选择D3单元格，在Excel菜单中选择【Crystal Ball】|【Copy】选项，复制D3单元格的参数设置。然后选择D4:D40单元格区域，单击【Paste】按钮将D3单元格中的参数设置全部复制到所选的单元格区域，如图1-42所示。

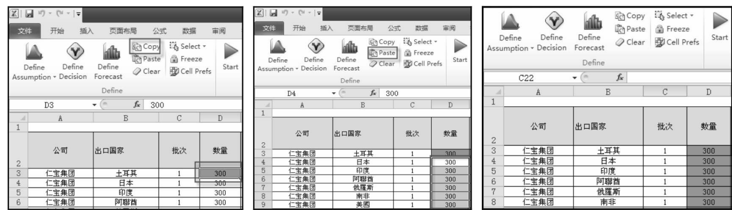


图1-42　快速复制参数

其他几个变量的复制步骤与上面相同，不再一一叙述。

如果认为某几个参数的设置应该与复制的内容不同，则可以单独修改，这样也可以对比不同参数的运行结果。

假设出口到日本时每次抽样的数量为3台～7台且抽检的概率为70%，则选择G4和J4单元格修改，操作如图1-43所示。

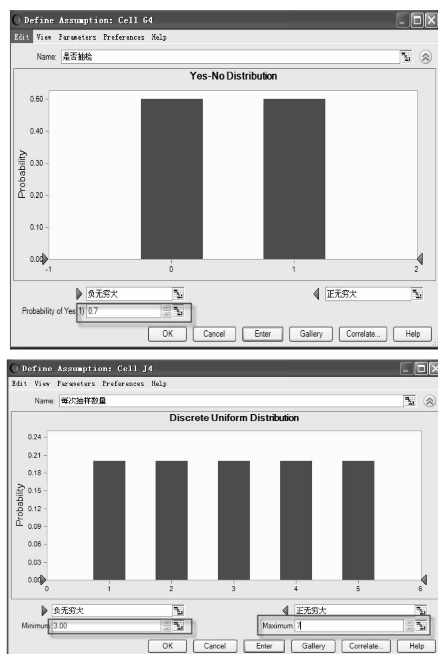


图1-43　修改概率分布示例

在【Yes-no】下拉列表框中设置检验的概率值70%，在【Discrete Uniform】微调框中设置最小值为3，最大值为7。

其他出口国家根据实际情况修改如下。

·印度：产品合格率的最大值为98%，抽检的概率为35%。

·阿联酋：每批次产品数量的均值为200，标准差为10；每次抽样数量为1台～4台。

其他国家的检验参数与土耳其保存一致。

（7）运行模型

选择Excel菜单中的【Crystal Ball】|【Start】选项，由于模型中有18个预测变量，所以会有18个计算结果。将参数不一样的几个国家的计算结果进行对比，为了对比查看方便，在预测变量概率图中选择【View】|【Statistics】选项，结果如图1-44所示。

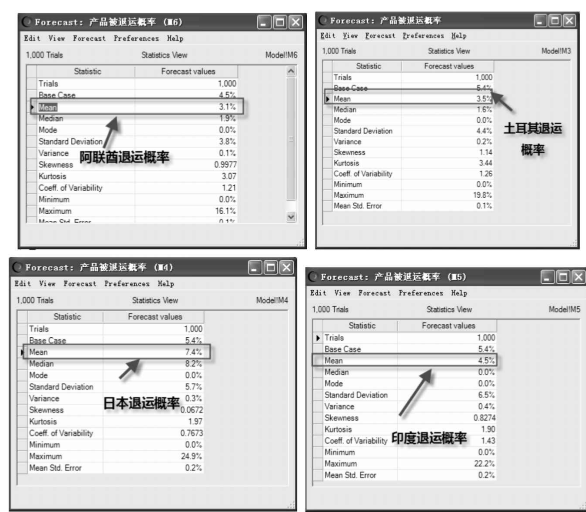


图1-44 出口货运退运概率计算结果

从以上分析可以看出在检验非常严格的假定国家出口货运被退运的概率明显高于检验宽松的国家。例如，日本的抽检的概率为70%，每次抽样3台~7台；阿联酋为每批次产品数量均值为200，标准差为10，每次抽样数量为1台~4台。以上4个国家的假设条件与分析结果如表1-2所示。

表**1-2** 各国家出口条件及退运概率值汇总

国家或地区	假设变量				预测变量（退运概率）		
	每次货物数量	产品质量合格率	抽检概率	每次抽样数量	最大值	均值	最小值
土耳其	均值为 300，标准差为 20 的正态分布	均值为 98%，标准差为 1%的正态分布且最大值为 99.9%	50%	1 台～5 台/次	19.8%	4.5%	0
日本	均值为 300，标准差为 20 的正态分布	均值为 98%，标准差为 1%的正态分布且最大值为 99.9%	70%	3 台～7 台/次	24.9%	7.4%	0
印度	均值为 200，标准差为 10 的正态分布	均值为 98%，标准差为 1%的正态分布且最大值为 98%	35%	1 台～5 台	22.2%	4.5%	0
阿联酋	均值为 300，标准差为 20 的正态分布	均值为 98%，标准差为 1%的正态分布且最大值为 99.9%	50%	1 台～4 台	16.1%	3.1%	0

（8）结果分析

从以上概率分析可以看出由于各国进口检验的力度不同，因而导致出口退运风险不同。根据假设条件的计算结果来看，出口到日本的风险最高，阿联酋的风险最低。出口企业也可以更加合理地来安排出口国家的产品质量，出口到日本的产品应该具有更高质量。

从长期的行为来看，出口到各个国家被退运的概率应该接近模型中的均值，如长期出口到阿联酋产品被退运的概率均值为3.1%。如果实际退运概率达到15%以上，则可以考虑该国是否有贸易保护的因素。

Miss Ju：“我现在明白为什么叫‘冒险记’了，不管你的产品是否合格，即使100%的合格都有可能被退运，是吗？”

Mr Shu：“是的，这其实也是概率在风险分析中的一个应用，我们可以定量计算出风险值是多少。”

Miss Ju：“真想不到概率分析还可以在进出口监管应用，看来减少退运概率的最正确的途径就是提高产品质量。”

Mr Shu：“对，打造高质量的产品是打开市场的最佳途径。Miss Ju，通过这两个例子你能不能简单说说对Minitab与Crystal Ball两个软件的认识啊？”

Miss Ju：“Minitab让我认识了模拟数据和概率，通过Crystal Ball我打破了常规思维，以前认为Excel中一个单元格只能代表一个数据，现在知道它也可以代表一组数据。还是有概率分布的数据，有概率就有风险，只是概率不同罢了。因此预测是建立在概率基础上计算出来的结果，假设变量的正确设置非常重要。”

Mr Shu：“漂亮，这么快就进入角色了。对Crystal Ball的假设变量的设置有几点需要注意，一是假设变量单元格必须是数值，而不能是公式或文本；二是假设单元格没有绝对的限制，一般来讲，你可以在一个单元格中定义少于1 000个假设变量、决策变量和预测变量；三是Crystal Ball软件中自带22种概率分布，如果这些概率分布均不能反映实际数据的分布，用户可选择自定义分布。”

Miss Ju：“对于概率还有两个问题，一是既然概率分布在计算中这么重要，那么我们拿到一组数据，如我的月度通话数据怎样才能知道它处于何种分布呢？二是概率好像分很多种吧？它们之间有什么不一样？又有什么联系呢？”

Mr Shu：“哇，提出来的问题很专业嘛，我们会一一讨论。在举例说明之前我们声明一个问题，即Minitab、Crystal Ball和Excel在数据分析上各有伯仲。我们在讲解时根据例子最适合哪个软件来说明，并不是其他软件不能实现。例如，电话套餐选择这个案例用Excel和Minitab都可以模拟，只是步骤烦琐罢了。”

Miss Ju：“那我们来看看如何来识别分布吧？”

Mr Shu：“通过Minitab和Crystal Ball都可以轻松识别分布。”

1.4　慧眼识分布

分析选择电话套餐后，Miss Ju决定自己动手分析自己的通话时间与长途占比这两个数据处于何种分布，以选择适合自己的套餐。她兴致勃勃地收集了60组数据，并将其整理在Excel表中，如表1-3所示。

表1-3　自2009年每月通话时间及长途占比数据

时　　间	2009-1-1	2009-2-1	2009-3-1	2009-4-1	2009-5-1	2009-6-1
通话时间	454	408	434	440	438	454
长途占比	0.18	0.21	0.17	0.24	0.18	0.17

Miss　Ju：“通过前面的分析，我知道了数据的分布对决策影响较大，我已收集了数据，怎么 知道它们处于何种分布？”

Mr Shu：“我们分别通过Minitab和Crystal Ball两款软件来分析，看看结果。”

在Minitab中的操作步骤如下。

（1）转置数据

在Excel中转置数据，结果如图1-45所示。

Minitab - 通话时间分布分析.MPJ - [工作表 1 ***]

文件(F)　编辑(E)　数据(A)　计算(C)　统计(S)　图形(G)

↓	C1-D	C2	C3	C4	C5
	时间	通话时间	长途占比		
1	2009-1-1	438.6	0.18		
2	2009-2-1	423.7	0.21		
3	2009-3-1	410.1	0.17		
4	2009-4-1	448.6	0.24		
5	2009-5-1	429.7	0.18		
6	2009-6-1	434.7	0.17		
7	2009-7-1	428.1	0.20		
8	2009-8-1	437.2	0.23		
9	2009-9-1	426.3	0.16		
10	2009-10-1	423.5	0.21		
11	2009-11-1	422.2	0.17		

图1-45　转置结果

（2）分布识别

打开通话时间分析.MPJ文件，在Minitab中选择【统计】|【质量工具】|【个体分布标识】选项，如图1-46所示。



图1-46　【个体分布标识】选项

（3）设置分布识别参数

显示【个体分布标识】对话框，选择【单列】单选按钮，选择“通话时间”变量。在【子组大小】文本框中输入1，选中【使用所有分布和变换】单选按钮，如图1-47所示。

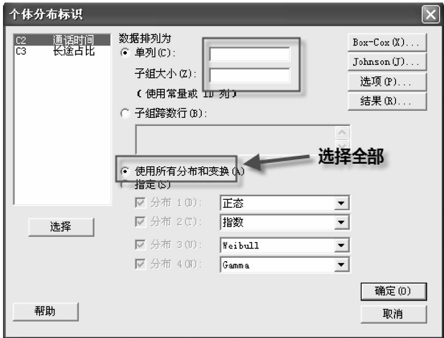


图1-47　设置参数

（4）解读分析结果

单击【确定】按钮，软件会自动对所有的分布和变换进行匹配计算，结果如图1-48所示。

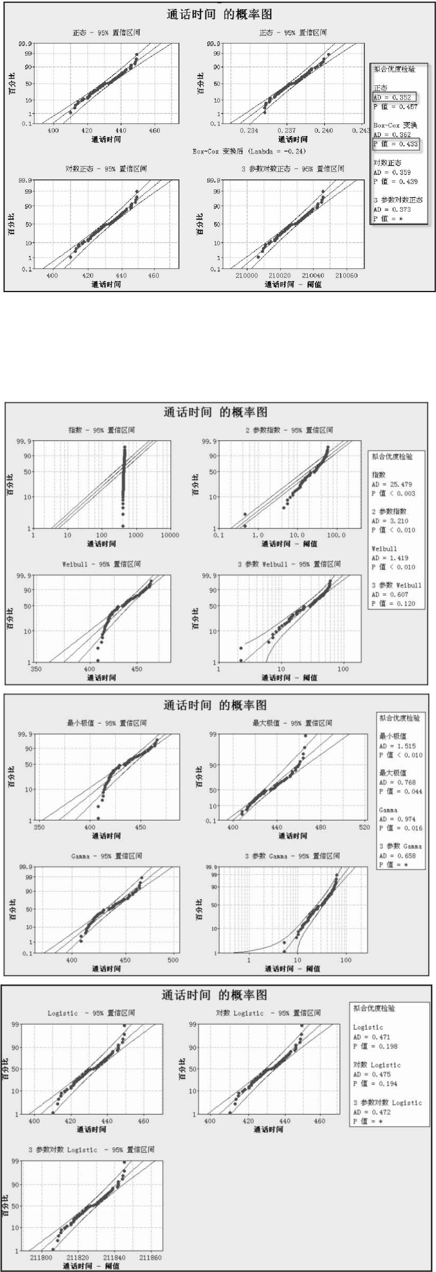


图1-48　计算结果

其中包括4个分析图形，有3个图形均有4个分布或变换的拟合图，一个图形有3个分布拟合图，一共15个图，即有15个分布和变换参与了拟合。

在每个图形的右侧有拟合优度检验值，分别是每种分布或变换的AD值和P值，它们是帮助我们选择最佳分布的关键参数。

同时在Minitab窗口中显示15种分布或变换的AD值与P值的汇总信息，如图1-49所示。

拟合优度检验

分布	AD	P	极大似然比 P
正态	0.352	0.457	
Box-Cox 变换	0.362	0.433	
对数正态	0.359	0.439	
3 参数对数正态	0.373	*	0.684
指数	26.239	<0.003	
2 参数指数	6.913	<0.010	0.000
Weibull	0.581	0.138	
3 参数 Weibull	0.343	0.434	0.042
最小极值	0.624	0.099	
最大极值	0.776	0.042	
Gamma	0.378	>0.250	
3 参数 Gamma	0.397	*	1.000
Logistic	0.471	0.198	
对数 Logistic	0.475	0.194	
3 参数对数 Logistic	0.472	*	0.753

分布参数的极大似然估计

分布	位置	形状	尺度	阈值
正态*	431.50115		10.35139	
Box-Cox 变换*	0.23747		0.00135	
对数正态*	6.06699		0.02403	
3 参数对数正态	12.25500		0.00005	-2.09597E+05
指数			431.50115	
2 参数指数			21.77332	409.72774
Weibull		46.88243	436.46513	
3 参数 Weibull		3.47803	35.07645	400.02273
最小极值	436.57041		9.27393	
最大极值	426.33669		9.78272	
Gamma		1763.43996	0.24469	
3 参数 Gamma		107.88322	0.99905	323.66377
Logistic	431.60731		6.12337	
对数 Logistic	6.06735		0.01420	
3 参数对数 Logistic	12.26353		0.00003	-2.11396E+05

* 尺度:调整后的极大似然估计

图1-49　Minitab中的分布计算汇总信息

选择一组数据服从何种分布首先看AD值，其值越小越好；其次看P值，P值最好大于0.10（设置值，也可以设置为0.05）；否则按照大小次序选择。

从以上分析结果来看，正态分布AD值最小，其值为0.352；P值为0.457。

该组数据的最佳分布拟合依次为正态分布＞　Box-Cox变换＞对数

正态分布＞3参数对数正态分布。

在汇总表中还罗列每种分布拟合出的参数，如正态分布的两个参数分别是“位置”为431，“尺度”为10.35。即均值为431，标准差为10.35。

AD值为Anderson-Darling统计量的简称，用来表征测量数据服从特定分布的程度。分布与数据拟合越好，此统计量越小。使用Anderson-Darling统计量可比较若干分布的拟合情况，以查看哪种是最佳分布，或者检验数据样本是否来自具有指定分布的总体。Anderson-Darling检验的假设为H_0，则数据服从指定分布；为H_1 则数据不服从指定分布。因为有假设检验，所以也有P值（在第4章中讨论假设检验）。

P值是用于检验AD的假设是否成立的概率值，如P＞0.05或者＞0.10，则可以认为AD的原假设是成立。但是Minitab并不始终为Anderson-Darling检验显示P值，因为某些情况下这个数值没有数学意义。在不存在数学意义的情况下根据AD值来判断，所以主要根据AD值的大小判别分布的拟合情况。

＊号表示的是如果不能直接计算出P值，则数据调整后的极大似然估算，P值无法计算时参考调整后的P值。

Johnson和Box-Cox变换分别是两种数据标准化的方法，有时可能会出现变化后数据的AD值最小的情况。但变换是将原数据进行标准化转换后拟合，不代表原始数据的分布，按照原始数据分布选择的优先次序时也暂不参考该次序。

（5）重复以上操作，确定“长途占比”分布

对“长途占比”进行分布拟合，计算结果如图1-50所示。

拟合优度检验

分布	AD	P	极大似然比 P
正态	1.151	<0.005	
Box-Cox 变换	1.151	<0.005	
对数正态	1.186	<0.005	
3 参数对数正态	1.203	*	0.802
指数	19.924	<0.003	
2 参数指数	3.065	<0.010	0.000
Weibull	1.209	<0.010	
3 参数 Weibull	1.347	<0.005	0.015
最小极值	1.389	<0.010	
最大极值	1.344	<0.010	
Gamma	1.199	<0.005	
3 参数 Gamma	1.370	*	0.489
Logistic	1.195	<0.005	
对数 Logistic	1.226	<0.005	
3 参数对数 Logistic	1.209	*	0.880
Johnson 变换	0.483	0.222	

分布参数的极大似然估计

分布	位置	形状	尺度	阈值
正态*	0.20850		0.03156	
Box-Cox 变换*	0.45532		0.03466	
对数正态*	-1.57923		0.15295	
3 参数对数正态	-1.11720		0.09550	-0.12019
指数			0.20850	
2 参数指数			0.05010	0.15840
Weibull		7.52490	0.22211	
3 参数 Weibull		1.80455	0.06258	0.15272
最小极值	0.22417		0.02861	
最大极值	0.19303		0.02765	
Gamma		43.96598	0.00474	
3 参数 Gamma		3.73627	0.01769	0.14241
Logistic	0.20822		0.01910	
对数 Logistic	-1.57658		0.09255	
3 参数对数 Logistic	-1.24672		0.06645	-0.08033
Johnson 变换*	0.03122		1.02118	

图1-50　“长途占比”分布拟合汇总计算结果

从以上分析看出AD最小值为正态分布，拟合的最佳程度排序为：正态分布＞Box-cox变换＞对数正态分布＞3参数对数正态分布。

Miss　Ju：“Mr　Shu，看来我的通话时间和长途时间都还比较固定嘛，都处于正态分布。”

Mr Shu：“从软件的分析结果来看是这样的。”

Miss　Ju：“我们前面在学习正态分布时知道位置参数为均值，尺度参数为标准差。它只有两个参数，分析出来的3个参数的那些分布，如3参数的Weibull分布的3个参数是哪些？”

Mr　Shu：“不错，现在学会举一反三了。等我们讲完Crystal Ball的操作再对分布进行一个总结，这样你就很清楚了。”

Miss　Ju：“好的，既然AD值是主要的选择参数，能不能在计算之后直接把AD值排序啊？那样不是一目了然吗？另外，Minitab中拟合了15种分布。好像也不是很全啊，像均匀分布和三

角形分布好像都没有拟合。”

Mr Shu：“是的，要是能排序就可以更加方便选择，每款软件的设置都不一样。在Crystal Ball中可以进行AD排序，拟合的分布相对Minitab也要多一些，该软件会自动将你的数据和每个连续分布概率分布进行匹配（注意只有连续性分布才能进行拟合）；同时显示每个分布的参数设置，能最好地描述数据特点。每次拟合的质量和优先度可以使用多个最佳拟合标准检验中的一个来判断，拟合排名最高的分布代表是最适合你的数据选择的分布。”

在Crystal Ball软件操作步骤如下。

（1）转置数据

在Excel中转置数据，结果如图1-51所示。

	A	B	C
1	时间	通话时间	长途占比
2	2009-1-1	454	0.18
3	2009-2-1	408	0.21
4	2009-3-1	434	0.17
5	2009-4-1	440	0.24
6	2009-5-1	438	0.18

图1-51　转置结果

（2）数据拟合

打开通话时间分布分析.xls文件，任意选择一个空单元格。在

Excel菜单中选择【Crystal Ball】|【Define Assumption】选项，选择【Distribution Gallery】选项，单击【Fit】按钮，如图1-52所示。

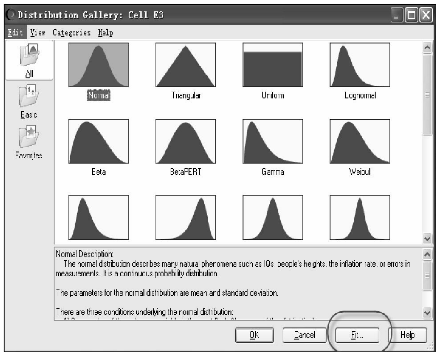


图1-52 【Fit】选项

选择B2:B61单元格，即对“通话时间”数据进行分布拟合，【Fit Distribution】对话框如图1-53所示。

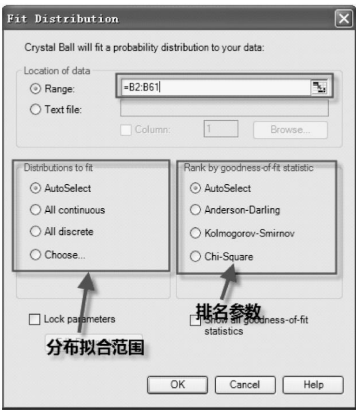


图1-53 【Fit Distribution】对话框

如图中所示设置有关参数。

在【Range】下拉列表框中选择B2:B61单元格区域。

【Text　file】单选按钮后下拉列表框用于选择外部文本，如果数据在一个单独的文本中，则选择浏览选择文件。

在【Distributions to fit】选项组中选择需要拟合的分布，可以自动匹配或手动选择，默认为自动选择。

在【Rank by goodness –of-fit statistic】选择拟合数据完成后使用的检验标准，Anderson-Darling、Kolmogorov-smirnov和Chi-square（卡方检验）分别为3种检验方式，选择自【Auto Select】单击按钮，默认按照AD值排序，系统默认AD值最小的分布拟合度最好。

（3）解读分析结果

单击【OK】按钮，计算结果如图1-54所示。

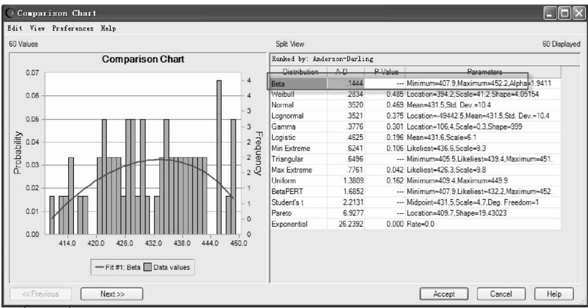


图1-54　分布拟合计算结果

从分析结果看，AD值最小的为Beta分布，拟合优度排序为：Beta分布＞Weibull分布＞正态分布。在Minitab中拟合时正态分布的AD值与Crystal　　Ball分析结果一致，均为0.352。在默认情况下，Crystal Ball在模拟拟合分布时的范围较Minitab大。但相同的分布拟合值一致，如本例中的正态分布和对数正态分布拟合。

Crystal　Ball拟合的正态分布对应的参数为Mean（位置）＝431.5

且Std.Dev（尺度）=10.4。

Minitab拟合出来的正态分布参数为Mean（位置）=431.5且Std.Dev（尺度）=10.4。

单击【Next】按钮可以查看不同分布的拟合情况，Crystal Ball已排序拟合出后的AD值和P值，并将拟合的参数放在图表的最右端，此时Beta分布的参数分别为Minimum=407.9、Maximum=452.2、Alpha=1.94114且Beta=1.70374。

（4）定义假设变量

单击【Accept】按钮，将之前选择的空单元格定义为一个假设变量，如图1-55所示。

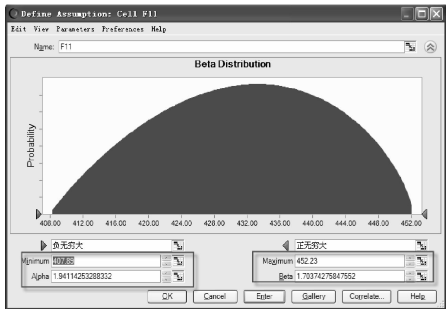


图1-55　定义假设变量

此时系统自动将拟合出来的最优分布的参数设置为初始参数，当然分布的相关参数可以修改。如不需要修改，单击【OK】按钮，所选单元格代表一组Beta分布的数据。

Miss　Ju："即只要有了假设变量的历史数据，我们在定义模型过程中可以将历史数据形成的分布直接定义在假设变量中。"

Mr Shu："没错，这样在建模过程中就可以一气呵成地完成假

设变量的概率分布定义。”

（5）模拟“长途占比”变量

重复以上步骤，“长途占比”变量的分析结果如图1-56所示。

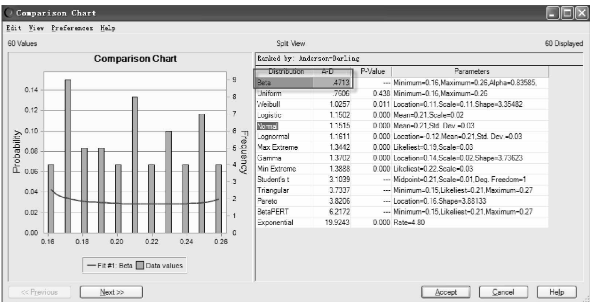


图1-56　“长途占比”变量的分析结果

与“通话时间”变量分析一样，Crystal Ball中拟合了Beta和Uniform分布。除去Beta和Uniform分布，其他拟合的排序为：正态分布对＞数正态分布＞Gamma分布，与Minitab分析结果基本相同。

选取二者拟合的正态分布的拟合优度值和参数进行对比，拟合结果一致。如表1-4所示。

表**1-4**　计算结果对比

软　　件	均　　值	标　准　差	**AD**值
Crystal Ball	0.21	0.03	1.151
Minitab	0.208	0.03	1.151

Miss Ju：“总体来讲，两款软件拟合分布的结果相差不大。”

Mr Shu："是的！"

Miss　Ju："要是遇到比较复杂的数据，前后两段时间的分布可能完全不一样，如前半年是正态分布；后半年是对数正态分布呢？这种数据我们直接用一种分布来表征就不准了，这时该怎么办？"

Mr　Shu："果然是越学越精了，在Crystal　Ball软件中有22种概率分布。如果这些分布还不能满足模型的要求，还可以自定义分布。"

1.5　分布72变

自定义分布——定义离散型分布

例如，某种商品的价格只有可能为5、6和8共3种。每种商品价格出现的机会均等，概率均为1/3，操作步骤如下。

（1）打开自定义窗口

在Excel中选择任一空单元格，在菜单栏中选择【Crystal　Ball】|【Define　Assumption】选项，选择【Distribution　Gallery】窗口中的【Custom】选项，如图1-57所示。

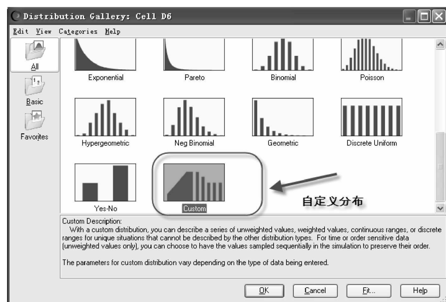


图1-57　【Custom】选项

（2）设置参数

单击【OK】按钮，选择【Parameters】|【Weighted Values】选项。在【Value】下拉列表框中输入3、6和8，在【Probablity】文本框中输入1/3，如图1-58所示。

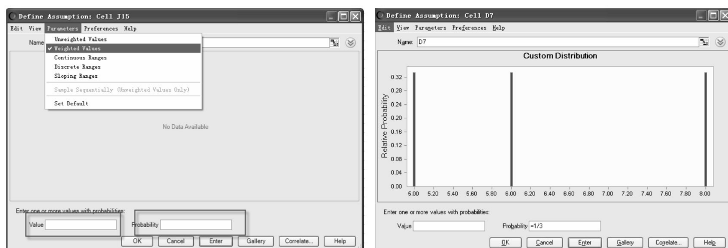


图1-58　设置参数

这样就形成了一个概率均为1/3的简单离散型分布。

（3）保存分布

单击【Edit】|【Add to Gallery】选项，如图1-59所示。

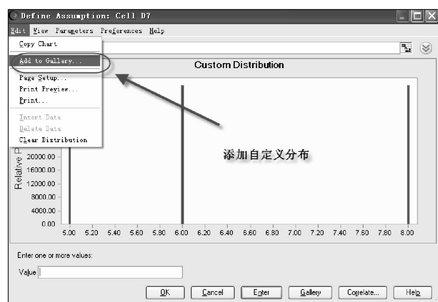


图1-59 【Add to Gallery】选项

显示【Add to Gallery】对话框，命名和分类分布，如图1-60所示。

选择【Favorites】选项，单击【OK】按钮，该分布保存在Favorites分布库中。

如果在输入数据时未输入概率值，则默认概率为1。但在运行计算时系统的总概率不会超过1，软件会自动将概率均分。



图1-60 命名和分类分布

自定义分布——删除自定义分布

删除自定义分布的操作步骤如下。

（1）选择需要删除的分布

选择任意Excel单元格，在Excel菜单中选择【Crystal Ball】|【Define Assumption】选项，双击选择需要删除的自定义分布，如图1-61所示。

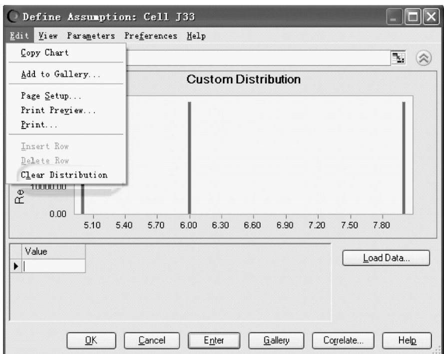


图1-61 【Define Assumption】选项

（2）删除自定义分布

选择【Edit】|【Clear Distribution】选项，或右击后选择【Clear Distribution】选项。

如果需要修改已定义的分布，选中后选择【Parameters】选项或右击后选择【Parameters】下级菜单中的分布类型，如图1-62所示。

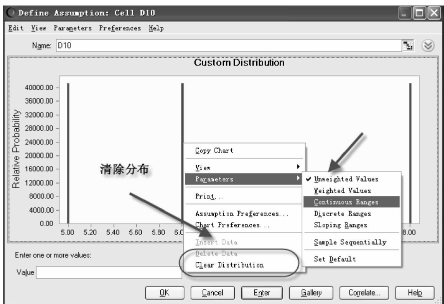


图1-62　修改自定义分布参数

自定义分布——定义连续性分布

【案例背景】

某超市为了更好地满足客户的购物需求，打算分析客户行为。他们从数据库中收集了大量的客户消费数据。分析后得知某商品需求量在1套／天～10套／天的频率为60%，12套／天～20套／天的频率为40%。将该频率数据设置为概率分布，将该概率分布自定义为连续性分布，操作步骤如下。

（1）打开自定义窗口

同定义离散型分布，选择【Define Distribution】|【Distribution Gallery】选项，选择【Custom】选项，操作过程略。

（2）设置参数

选择【Parameters】|【Continuous Ranges】选项，输入相关参数，如图1-63所示。

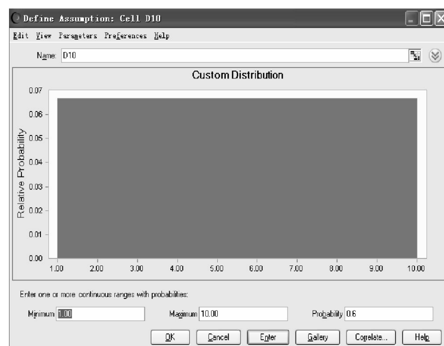


图1-63　输入参数

输入的参数如表1-5所示。

表**1-5**　输入的参数

Minimum	Maximum	Probablity
1	10	0.6
12	20	0.4

输入完成后的概率分布如图1-64所示。

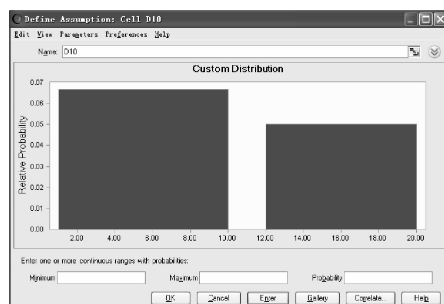


图1-64　概率分布

（3）添加连续性分布

同增加离散型分布，在Crystal　Ball中选择【Edit】|【Add　to Gallery】选项添加分布。

除了可以单独自定义连续性分布和离散分布之外，Crystal　Ball还可以定义既有连续性分布，又有离散型分布的组合分布。

自定义分布——定义组合分布

当发现某组数据是既有连续性分布，又有离散型分布时，可以定

义组合分布。例如，在上例中客户如果购买商品的数量在12套／天～20套／天之间时，不再是一种连续性分布，而有可能是一种步长为2的离散型分布。也就说这种商品的购买量在12套／天～20套／天之间时，只可能会出现12、14、16、18和20套这几种可能性。其他购买量，如13不会出现（这是一种假设，为了理解分布的需要，在实际生活中不太可能出现这种情况）。我们如果此时不重新定义分布，则只需要修改该上例中的自定义分布，操作步骤如下。

（1）选择分布

选择任意单元格，在Excel菜单中选择【Crystal Ball】选项。选择【Define Assumption】选项，打开上例中的自定义分布。单击12套／天～20套／天之间的分布，该分布底纹变成浅绿色，如图1-65所示。

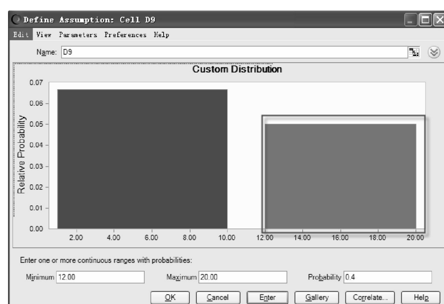


图1-65　选择需要修正的分布

（2）修改属性

选择【Parameters】|【Discrete　Ranges】选项，设置Min为12，Max为20，Pro为0.4，Step为2，则以上分布变为组合分布，修改后的组合分布如图1-66所示。

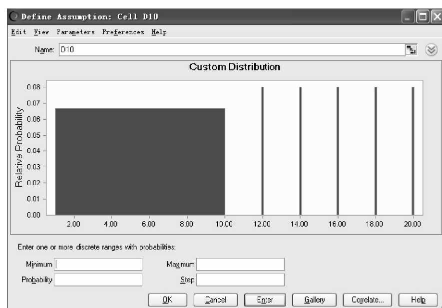


图1-66　修改后的组合分布

单击【show　more】按钮可以更方便地设置多组数据，如图1-67所示。

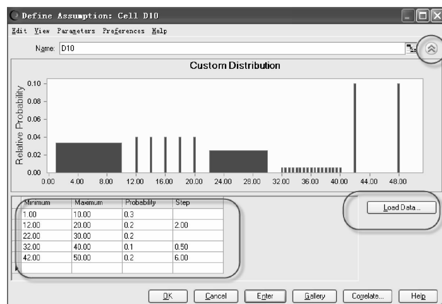


图1-67　设置多组数据



图1-68　【Load Data】对话框

单击【Load　　　Data】按钮，直接将Excel表格中的数据加载至Crystal Ball中，【Load Data】对话框如图1-68所示。

自定义分布——其他功能

在概率分布定义中还有一项功能可约束变量的范围，在前面的海关检测案例中我们已提及过。即可以修正一些不符合常理或逻辑的范围，如产品的合格率不可能超过100%等。

例如，某商品价格历史数据服从均值为10且标准差为1的正态分布；未来该商品可能会涨价，但对未来的销售价格预测不会超过12，即价格的分布中超出12的均没有意义，因此在定义假设变量时可以对该分布进行截断。操作步骤如下。

（1）定义假设变量

在Excel中选择任意空单元格，选择【Crystal Ball】|【Define Assumption】选项，选择【Distribution Gallery】|【Normal】选项。输入均值为10，标准差为1，如图1-69所示。

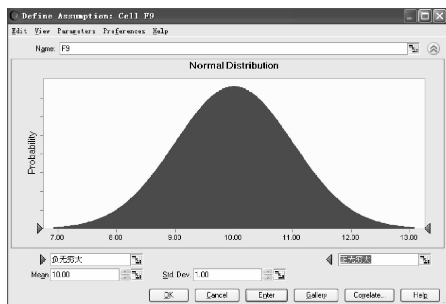


图1-69　输入值

（2）设置假设变量范围

单击【Show more】按钮，在右下角下拉列表框中输入12，或者直接向左拉动小三角到12， 如图1-70所示。

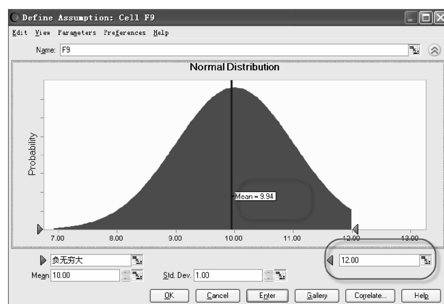


图1-70　设置假设变量范围

从以上设置看出此时的分布不再是一个完整的分布，分布截断后参数与完整的分布已有所不同，如上图中的均值已经由截断前的10变为9.94。

Miss　Ju：“真的是太过瘾了，我们从足球预测开始聊到概率分布。了解了Minitab和Crystal Ball两款软件在概率分布上的一些应用，还对比了两款软件的异同点。”

Mr　Shu：“是的，我们在上面聊到Crystal　Ball用在风险分析和蒙特卡洛建模的一些基本操作，以及定义假设变量和预测变量，还可以拟合分布和自定义分布。Minitab在计算机仿真和概率图形上的一些应用等都比较实用。”

Miss　Ju：“看来正确地识别分布真的太重要了，我们一直在谈概率的一些知识点，概率不是主要应用在风险分析和预测上面吗？什么时候可以开始学习预测的一些应用。”

Mr　Shu：“不要着急，我们在第1章的后面专门讨论预测的一些应用。预测也是建立在概率的基础上的，不是吗？在总结分布之前对Crystal Ball软件中的概率拟合还有没有疑问？”

Miss　Ju：“我注意到在定义假设变量时不仅可以通过Fit来拟

合历史数据的分布，好像还有一个【Correlate】按钮，这个单词是相关的意思，它是用来查找两组数据之间的相关性吗？”

Mr Shu：“观察力很强！这也是优秀数据分析师必备的素质之一。Correlate的确是相关的意思，但是在Crystal Ball中不仅仅用来分析数据之间相关性，分析数据相关性在Minitab中也可以方便地进行，还可以定义两个假设变量之间的相关性，这样会让风险分析更加准确。”

Miss Ju：“有点晕，假设变量本来就是一组概率分布数据，两个假设变量之间定义相关性又有什么作用呢？”

Mr Shu：“举个简单例子吧，我们前面提到的电话套餐中有两个假设变量，即通话时长和长途占比，我们分别对这两个变量进行了概率分布假设，或从历史数据进行分布拟合后定义。打电话时间越长和长途电话占比重大的概率值比较高，即长途电话和通话总时长总存在一点相关性，不可能是完全独立的两个变量。”

Miss Ju：“那么怎么操作？添加相关性之后的数据模型计算出来的结果会有什么不一样吗？”

Mr Shu：“还是通过前面的案例来分析，即可知道计算结果会有哪些不一样。”

假设变量对对碰

在上面的学习中我们已知道如何自定义假设变量的分布，其中每个假设变量均是独立的，没有相关性。但在实际生产中假设变量往往也存在一定相关性，此时也可以在模型中设置假设变量之间的相关性

后运算。

继续以手机套餐选择为例，该例中手机“通话时长”和“长途占比”为两个独立的假设变量。但是往往长途电话越多，通话时间越长，二者之间可能存在一定程度的相关性。操作步骤如下。

（1）选择第1个假设变量

打开Cell Phone-correlate.xls工作表，此时B9和B10单元格已设置为假设变量。选择B10单元格，单击【Define Assumption】|【Distribution Gallery】选项，选择假设变量“长途占比”，如图1-71所示。

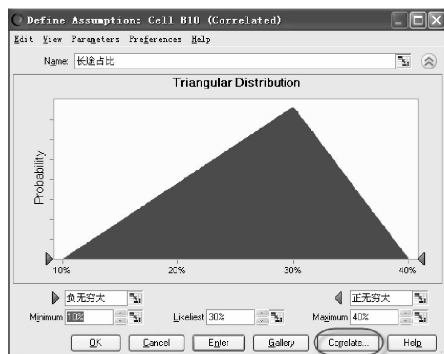


图1-71　选择假设变量“长途占比”

（2）选择另一个假设变量

单击【Correlate】按钮，单击【Choose】按钮选择另一个变量“通话时长”（此时它仍然是均值为400，标准差为20的正态分布），如图1-72所示。

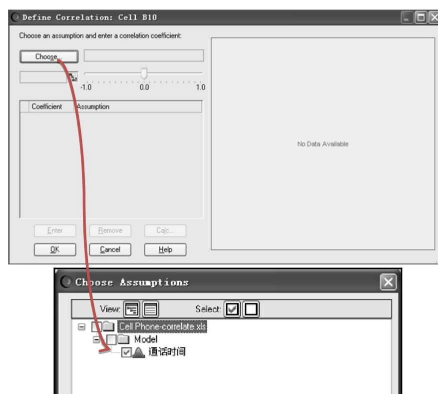


图1-72　选择另一个假设变量

（3）设置相关系数

在不清楚相关系数为多少的情况下，假设“长途占比”与“通话时长”两个变量之间的相关系数为0.7（-1为完全负相关，1为完全正相关）。设置完成后如图1-73所示。

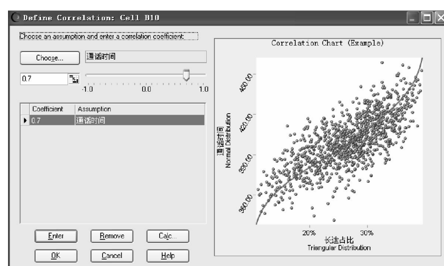


图1-73　设置相关系数

（4）设置前后对比

单击【OK】按钮，重新对该模型进行模拟运算后预测变量“话费之差”结算结果分布图如1-74所示。

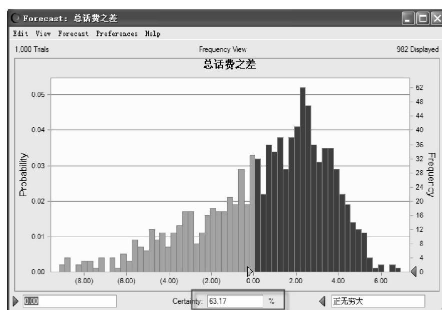


图1-74　预测变量“话费之差”结算结果分布

未设置相关性之前的预测变量“话费之差”分布如图1-75所示。

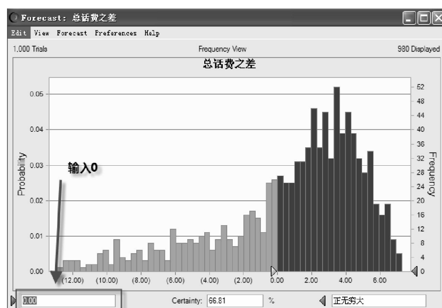


图1-75　未设置相关性之前的预测变量“话费之差”分布

设置相关性后“通话时长”与“长途占比”两个假设变量的敏感性分析结果对比如图1-76所示。

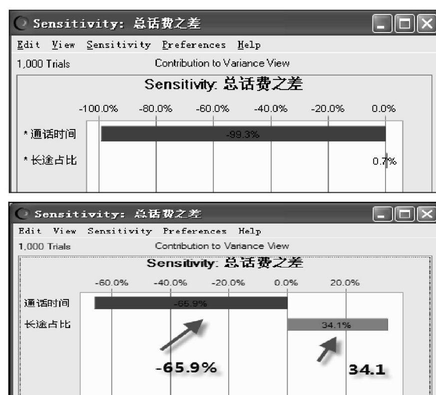


图1-76　敏感性分析结果对比

从以上结算结果可以看出设置相关性后预测变量“总话费之差”大于0的概率较之前有所降低，说明在设置相关性之后不确定性更多。所以模型中的预测变量在计算过程中更加谨慎，计算预测变量大于0的概率值会低一些。

敏感性分析也有很大变化，设置相关性之后假设变量“通话时长”占99.3%。这是因为二者之间有较强的相关性，所以敏感性通过其中一个变量来体现。

（5）验证二者之间的相关系数

相关系数0.7是在不清楚真实相关系数的情况下给定的假设系数，真正的相关系数又是多少？在Excel工作表中的F列和G列分别为一段时间的通话时长和长途占比的历史数据，可通过Crystal Ball软件中自动求出两个变量之间的相关系数。

在第（3）步设置相关系数界面时单击【Calc】按钮，在【First cell range】下拉列表框中输入单元格F2:F642区域，在【Second cell range】下拉列表框中输入单元格G2:G642区域，如图1-77所示。

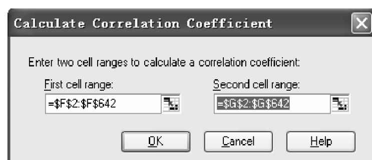


图1-77　输入区域

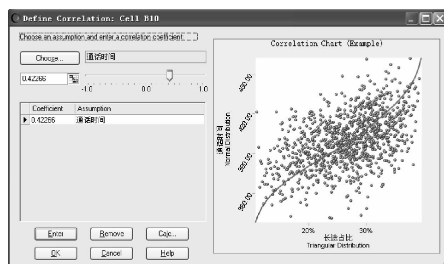


图1-78　计算两组数据相关性

单击【OK】按钮，系统就会自动将计算出来的相关系数0.4226代入数据模型中参与计算，具体操作这里不再叙述。

Miss　Ju：“原来假设变量之间还可以设置相关性并把它应用到模型中，这样模型与实际是不是更加吻合？”

Mr Shu：“没错，如果想让模型与实际更加吻合，除了假设变量与预测变量之外我们还可以加上约束条件和决策变量，这样才是真正一个完整的模型。例如，电话套餐这个模型还没有告诉我们答案。即究竟影响选择何种套餐，也没有在一定的约束条件下计算。”

Miss Ju：“听起来好像还有很多知识，接着来举例说明吧。”

Mr　Shu：“不要着急，我们本章的内容重点是理解风险与概率，有概率就存在风险。在第2章我们会涉及决策变量和约束条件

的内容。我们上面只重点说明了正态分布，将其他分布也串联起来吧。”

概率分布串串香

在现实世界的现象中人们必须考虑到随机性，即我们所关注的量可能是随时变化的，称为“随机变量”。例如，每次彩票的中奖号码和生男生女的可能性等。

随机变量概率也是概率与数理统计中最重要的概率，按照数据的类型可分为离散型变量和连续性变量。例如，产品“合格”与“不合格”，以及服务满意度的“满意”与“不满意”等为离散型变量，也称为“计数型数据”；蒸汽管线中的蒸汽流量、车床车间螺钉的长度和产品中某种组分的含量等均为连续性数据，也称为“计量型数据”。

在观察一种概率分布密度图形时，主要观察形状、散布和中心位置。形状参数主要是表征分布图形偏态还是正态，主要的形状参数有偏度和峰度等；散布参数表征图形的集中程度，是宽还是窄，主要有R （极差）、方差和标准差等；中心位置参数主要表征数据的集中程度，主要有均值、中位数和众数等。

在Minitab等统计学软件中将这些参数分别称之为“形状参数”、“尺度参数”和“位置参数”，大部分概率分布只需要其中两个参数；另外一个参数是固定值。还有部分概率分布需要“域参数”，即开始位置参数。它指定X 轴从哪个位置开始，该参数不参与数学期望及方差的计算。绝大多数概率分布图主要需要“位置”和“尺度”两个参数，其组合不同的概率分布函数形成不同的概率分布图形。在串联概率分布之前首先理解如下重要的统计学名词。

1. 数学期望

随机变量X的数学期望反映了X的平均取值，记$E(X)$。

（1）当X是离散型数据时$E(X)=\mu=\Sigma x_n Tn$。

假设进行掷骰子试验，试验次数为200次并将每次骰子的点数记录下来，数据如表1-6所示。

表**1-6**　掷骰子试验数据

点数	1	2	3	4	5	6	总计
次数	40	30	25	50	25	30	200
比率	0.2	0.15	0.125	0.25	0.125	0.15	1

如想求得每次骰子的平均值，我们在考虑次数的情况下一般计算方法为：

$$\mu=(40\times1+30\times2+3\times25+4\times50+5\times25+6\times30)/200=3.4$$

如果将分母200分别放入括号中的每一项，即先计算出每种点位所占的比率p值，平均值的计算可等价换成如下公式：

$$\mu=(0.2\times1+0.15\times2+3\times0.125+4\times0.25+5\times0.125+6\times0.15)=3.4$$

根据第2种计算方法可以得出一般的离散型随机变量数学期望$E(X)$的计算方法：

$$E(X)=\mu=x_1p_1+x_2p_2+x_3p_3+......+x_np_1=\Sigma x_nP_n$$

（2）当X是连续性数据时$E(X)=\mu=\int_{-\infty}^{\infty}xp(x)\mathrm{d}x$。

一个连续随机变量的概率密度函数有如下两个重要特征。

（1）位于概率密度函数曲线下的面积等于1。

（2）随机变量的两个给定值a 和b 之间的概率等于位于a 和b 之间曲线下的面积，如Z 值分布在1.5～0.8之间的概率为图示阴影面积，如图1-79所示。

正态分布是最重要的连续分布，其概率密度函数形成的图形如钟，如图1-79所示。

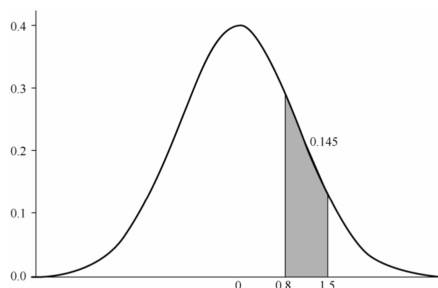


图1-79　概率之间面积

假设分布密度用p （x ）来表示，按照微积分理论，曲线下的面积为1。即$A=\int_{-\infty}^{\infty}p(x)\mathrm{d}x=1$ ，根据理论推算，可以计算出连续性分布概率的均值为：

$$E(X)=\mu=\int_{-\infty}^{\infty}xp(x)\mathrm{d}x$$

对于应用统计学来讲，我们不必深究计算E （X ）的过程。重要的是理解均值概念，熟悉几种常用的连续性分布的公式即可。

2．方差

均值即数学期望反映的数据中心点，但这对了解一组随机变量的波动情况还不够。相同均值的两组数据的波动情况可能会差异较大，

方差反映了随机变量取值相对于其数学期望的波动程度，即分散程度，记为D（X）。对于任一个随机变量，方差计算的公式为：

$$D(X)=E(X^2)-[E(X)]^2$$

离散型分布与连续性分布在计算D (X)会有所不同，在随后介绍各分布时会说明。与理解数学期望一样，我们不必深究复杂的理论推导过程，熟悉并记住几种常用的方差公式即可。

3．参数

大部分概率分布函数需要“形状”和“尺度”参数进行计算，下面详细介绍。

正态分布与卡方分布

概率分布从最重要的正态分布开始，其分布名称为“正态分布”。位置参数为均值，尺度参数为方差。

n 个服从标准正态分布的随机变量的平方和构成一新的随机变量，其分布规律称为χ^2 （k ）分布，其中参数n 称为“自由度”。卡方值都是正值，卡方分布是一个正偏态分布。不同的自由度决定不同的卡方分布，自由度越小，分布越偏斜。

为了进一步了解卡方分布与正态分布之间的关系，通过Excel数据拟合说明，操作步骤如下。

（1）拟合5组正态分布数据

新建Excel工作表，命名“分布模拟”。利用Normdist函数模拟5组均值为0且标准差为1的标准正态分布的数据。每组数据100行分别存

放在5列中，模拟完成后如图1-80所示（省略后80组数据）。

5组服从标准正态分布的数据				
x1	x2	x3	x4	x5
0. 211986	-1. 36755	0. 275274	1. 104545	-0. 61439
1. 408232	-1. 3855	-0. 21794	-1. 55095	1. 015335
0. 503196	1. 231635	-0. 43084	-0. 17768	1. 176827
-0. 1293	0. 77795	-0. 25546	0. 961109	-0. 17488
-0. 80305	-0. 87753	-2. 48376	-0. 0435	-1. 05015
0. 227425	0. 740899	1. 118592	0. 566577	-1. 18112
0. 474831	-0. 01266	1. 023825	-1. 24597	0. 278058
0. 080909	-1. 45867	0. 91328	0. 036842	-1. 34415
0. 393114	-0. 06787	-0. 73537	-1. 81027	-0. 06227
0. 318771	0. 979112	-0. 00983	-0. 18258	2. 045836
-0. 92824	0. 113653	1. 277858	-1. 80949	-0. 05032
-0. 8177	0. 669193	-1. 00895	0. 609778	0. 87472
-0. 54311	-1. 23147	-0. 55681	1. 333096	1. 163626
-0. 50719	-1. 50261	1. 857106	-1. 32846	-1. 85668
0. 094808	-0. 2908	2. 331726	-1. 16288	0. 134005
-0. 60453	-0. 58046	-1. 91174	0. 070552	0. 554494
0. 16053	1. 239844	-0. 44228	-1. 54034	-0. 64085
-0. 40446	-0. 25744	0. 013043	1. 682502	0. 910036

图1-80　5组正态分布数据

（2）对数据进行平方

在F3单元格中输入公式“＝A3×A3”，将A列数据进行平方。其他4列依次操作，结果数据如图1-81所示。

各列进行平方					自由度为的3的卡方分布	自由度为的5的卡方分布
x1^2	x2^2	x3^2	x4^2	x5^2	chi1^2	chi2^2
0. 044938	1. 870194	0. 075776	1. 22002	0. 377476	1. 990908108	3. 588403869
1. 983117	1. 919621	0. 047496	2. 405439	1. 030905	3. 950234486	7. 386578076
0. 253206	1. 516926	0. 185622	0. 031571	1. 384922	1. 955754486	3. 372246931
0. 016718	0. 605206	0. 065262	0. 92373	0. 030584	0. 687186296	1. 641500072
0. 644882	0. 770056	6. 16905	0. 001893	1. 102816	7. 583987529	8. 688695794
0. 051722	0. 548932	1. 251248	0. 321009	1. 395039	1. 851902213	3. 567949905
0. 225465	0. 00016	1. 048218	1. 552445	0. 077316	1. 273842657	2. 903603288
0. 006546	2. 127706	0. 834081	0. 001357	1. 806743	2. 968333327	4. 776433453
0. 154539	0. 004606	0. 540775	3. 277084	0. 003878	0. 699920109	3. 980882247
0. 101615	0. 95866	9. 66E-05	0. 033335	4. 185443	1. 060371892	5. 279150094
0. 861633	0. 012917	1. 632922	3. 274237	0. 002532	2. 507471222	5. 784239847
0. 66864	0. 447819	1. 017985	0. 371829	0. 765135	2. 134443566	3. 271407599
0. 294968	1. 516528	0. 310042	1. 777146	1. 354025	2. 12153809	5. 252708916
0. 257246	2. 257845	3. 448844	1. 764801	3. 447256	5. 963934724	11. 17599174
0. 008989	0. 084566	5. 436947	1. 352279	0. 017957	5. 530501041	6. 900737513
0. 365462	0. 336939	3. 654749	0. 004978	0. 307463	4. 357149377	4. 669590425
0. 02577	1. 537212	0. 195614	2. 372642	0. 410683	1. 758596261	4. 541922145
0. 163589	0. 066276	0. 00017	2. 830811	0. 828165	0. 230035305	3. 889011966

图1-81　平方后的数据

（3）平方相加

将平方后的前3列相加，得到自由度，即n 值为3的卡方分布数据；将平方后的5列相加得到自由度为5的卡方分布数据。

卡方分布的期望和方差

卡方分布的均值为自由度n ，记为$E(\chi^2)=n$ 。卡方分布的方差为两倍的自由度（$2n$ ），记为$D(\chi^2)=2n$ 。在Minitab中只需要指定自由度n 即可确定数学期望（位置参数）和方差（尺度参数）。

卡方分布与F分布

F分布是以统计学家R.A.Fi三her姓氏的第1个字母命名的，也是一种连续性分布。

设X_1 服从以自由度为m 的卡方分布，X_2 服从以自由度为n 的卡方分布且X_1 与X_2 独立，则$F=(X_1/m)/(X_2/n)$的分布就是自由度为m 与n 的F分布。

在上面例子中K列和L列分别是自由度为3和5的卡方分布，将这两列相除可以得到分子自由度为3且分母自由度为5的F分布。操作过程略，结果如图1-82所示。

K	L	P
自由度为的3的卡方分布	自由度为的5的卡方分布	F分布
chi1^2	chi2^2	(chi1^2/3)/(chi2^2/5)
1.990908108	3.588403869	0.924695297
3.950234486	7.386578076	0.891309085
1.955754486	3.372246931	0.966593158
0.687186296	1.641500072	0.697721866
7.583987529	8.688695794	1.454761395
1.851902213	3.567949905	0.865063628
1.273842657	2.903603288	0.731184974
2.968333327	4.776433453	1.035756545
0.699920109	3.980882247	0.293033916
1.060371892	5.279150094	0.334767236

图1-82　F分布与卡方分布结果

F分布是一种非对称分布，它有两个自由度，即m 和n ，相应的分布记为F（m ,n ），m 通常称为“分子自由度”，n 通常称为“分母自由度”，不同的自由度决定了其形状。

F分布是一个以自由度m 和n 为参数的分布族。

正态分布与*t* 分布

t 分布又称“Student t分布”，其推导由英国人威廉·戈塞特（Willams·Gosset）于1908年首先发表，当时他还在爱尔兰都柏林的吉尼斯（Guinness）啤酒酿酒厂工作。虽然酒厂禁止员工发表一切与酿酒研究有关的成果，但允许他在不提到酿酒的前提下以笔名发表t分布的发现，所以论文使用了“学生”（Student）这一笔名。

我们知道服从正态分布的随机变量由两个参数来表征，即均值μ（总体均值）和标准差σ （总体标准差）。由于在实际工作中往往总体的标准差σ 是未知的，所以常用样本表征差X 作为σ 的估计值。为了与U变换区分，称为“t变换”。统计量t 值的分布称为“t 分布”，可以认为t 分布是总体标准差未知情况下的正态分布。

设随机变量X_1 和X_2 独立且X_1 服从标准正态分布，X_2 是服从自由度为n 的卡方分布（上面已经过推导），则$t = X_1 / t = X_1 \Big/ \sqrt{(\chi^2 / n)}$ 的分布就是自由度为n 的t 分布。

操作步骤是：将L列的卡方分布数据与自由度5相除，并放置在M 列中。将M 列中的数据开根号，数据放置在N 列中。A 列正态分布数据与N 列的数据相除即得到t 分布数据，在Excel模拟数据中模拟数据如图1-83所示。

L	M	N	O
自由度为的5的卡方分布			T分布
chi2^2	chi^2/n	开根号	x1/开根号
3.588403869	0.717681	0.84716	0.2502314
7.386578076	1.477316	1.215449	1.1586108
3.372246931	0.674449	0.821249	0.6127208
1.641500072	0.3283	0.572975	-0.225661
8.688695794	1.737739	1.318233	-0.609183
3.567949905	0.71359	0.844743	0.2692238
2.903603288	0.580721	0.76205	0.623097

图1-83　*t* 分布数据拟合

t 分布的特征是以0为中心，左右对称的单峰分布。*t* 分布是一簇曲线，其形态变化与*n*（自由度）大小有关。*n* 越小，*t* 分布曲线越低平；*n* 越大，*t* 分布曲线越接近标准正态分布（*μ* 分布）曲线。

正态分布与Gamma分布

与正态分布不同，Gamma分布是一种偏态连续分布，是统计学的一种连续几率函数。其参数*α* 称为“尺度参数”（Scale parameter），*β* 称为“形状参数”（Shape parameter）。当*β* 趋向于较大数值时，近似于正态分布。在统计学软件如Mintab中Gamma分布还需要一个域参数。该参数指分布开始的起始位置，一般情况下默认为0。在Minitab中模拟两种不同参数组合下的Gamma分布，操作步骤如下。

（1）打开设置窗口

在Minitab菜单中选择【图形】选项，选择【单一视图】选项。选择【Gamma】选项，如图1-84所示。

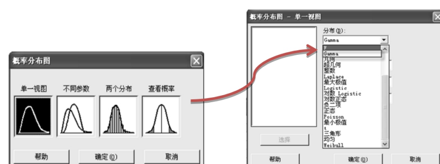


图1-84　选择【Gamma】选项

（2）参数设置

在弹出的对话框中设置形状参数为10，尺度参数为1，结果如图1-85所示。

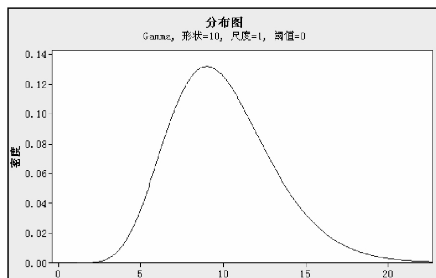


图1-85　Gamma分布示例结果

Gamma分布是一种重要的连续性分布，它和其他分布相关联，如指数分布、Pascal分布、Erlang分布、泊松分布和卡方分布。

Gamma分布与卡方分布

通过前面的分析，我们知道卡方分布是n（自由度）组正态分布的平方和；同时卡方分布与Gamma分布也有联系。

卡方分布实际就是尺度参数$\alpha=2$，形状参数$\beta=n/2$的一种Gamma分布。为了进一步了解二者之间的关系，在Minitab中对比分析，操作步骤如下。

（1）打开设置窗口

在Minitab菜单中选择【图形】选项，选择【两个分布】选项，分

别选择Gamma分布和卡方分布，如图1-86所示。



图1-86　选择两种分布

（2）设置参数

设置卡方分布与Gamma分布的参数，如表1-7所示。

表**1-7**　设置的参数

分　　布	自由度	尺度参数	形状参数
Gamma分布	不设置	默认为2	$n/2=5$
卡方分布	10	不设置	不设置

单击【多图形】按钮，选择【在同一图表的单独组块中】选项，结果如图1-87所示。

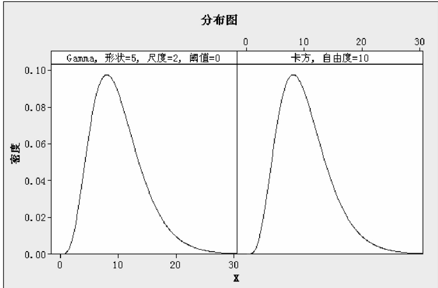


图1-87　Gamma与卡方分布结果

二者的分布图形相同。

Gamma分布与指数分布

当形状参数$\beta=1$且$\Gamma(1,\alpha)$时的Gamma分布就变化成为参数为α的指数分布，即指数分布是形状参数为1的Gamma分布。在Minitab中进行对比操作的步骤如下。

（1）打开设置窗口

在Minitab菜单中选择【图形】选项，选择【两个分布】选项，分别选择Gamma分布和指数分布，如图1-88所示。

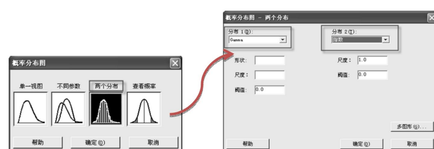


图1-88　选择两个分布

（2）设置参数

参数设置如表1-8所示。

表**1-8**　**Gamma**与指数分布参数设置

分　　布	形状参数	尺度参数	阈值参数
Gamma分布	1	10	0
指数分布	默认为1	10	0

单击【多图形】按钮，选择【在同一图表的单独组块中】选项，结果如图1-89所示。

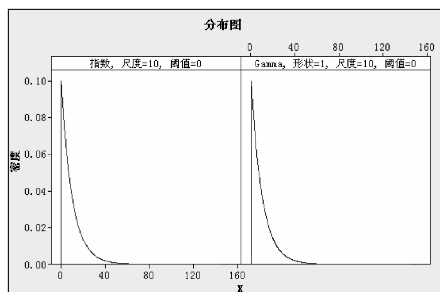


图1-89　Gamma和指数分布结果

二者图形完全一致，指数分布的尺度参数为λ　，其数学期望及方差分别为$\frac{1}{\lambda}$ 和$\frac{1}{\lambda^2}$ 。

指数分布与Weibull分布

Weibull分布广泛应用在寿命试验和可靠性试验中，它是瑞典科学家威布尔于1939年为描述材料强度而发现的一种分布，现以其名字命名。该分布函数有两个参数，即α　和β　。在Minitab中模拟α　＝1，β　＝2的Weibull分布和β ＝2的指数分布，操作步骤如下。

（1）打开设置窗口

在Minitab菜单中选择【图形】选项，选择【两个分布】选项，分别选择Weibull分布和指数分布，如图1-90所示。

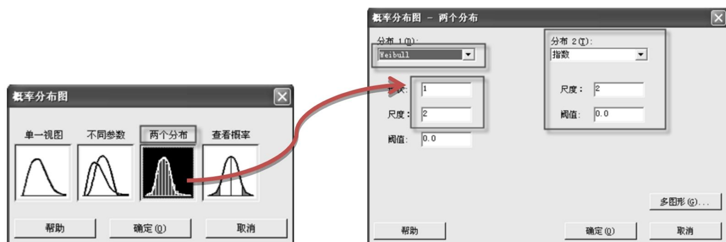


图1-90　两个分布对比

（2）设置参数

参数设置如表1-9所示。

表1-9　Weibull与指数分布参数设置

分　　布	形状参数	尺度参数	阈值参数
Weibull分布	1	2	0
指数分布	默认为1	2	0

单击【多图形】按钮，选择【在同一图表的单独组块中】选项，结果如图1-91所示。

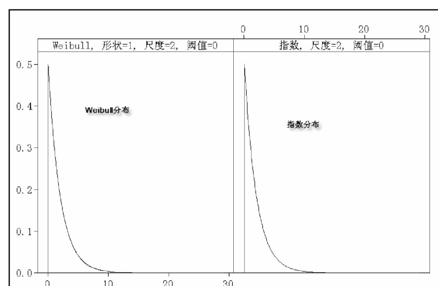


图1-91　Weibull和指数分布结果

显然，当Weibull分布＝2且β　　　＝1时与指数分布＝1时一样。Weibull分布在β　＝1的情况下只要α　值与指数分布一样，则分布形成的曲线相同，即Weibull分布是形状参数＝1的指数分布。

既然Weibull分布及Gamma分布与指数分布有关，由此我们可以进一步推断Weibull与Gamma分布对比且二者的形状参数和尺度参数相等时且形状参数都等于1时，这两个分布形成的曲线相同。通过Minitab拟合α　＝5且β　＝1的Weibull分布和α　＝5且β　＝1的Gamma分

布，操作步骤如下。

（1）打开设置窗口。

（2）设置参数，如表1-10所示。

表1-10　Weibull与Gamma分布参数

分　　布	形状参数	尺度参数	阈值参数
Weibull分布	1	5	0
Gamma分布	1	5	0

设置完成后两种分布的对比如图1-92所示。

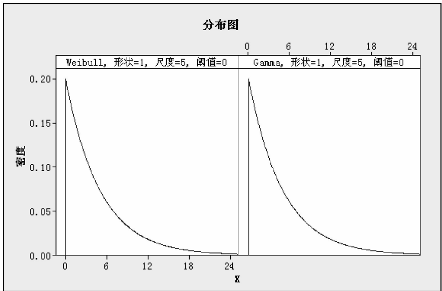


图1-92　Gamma及Weibull分布

Weibull分布的数学期望和方差计算公式为：

$$E(X)=a^{-\frac{1}{b}}\Gamma(1+\frac{1}{b})$$

$$D(X)=a^{-\frac{1}{b}}[\Gamma(1+\frac{1}{b})-\Gamma^{2}(1+\frac{2}{b})]$$

概率分布小结

Miss　Ju：“这有点头晕，被这些概率搞得有点乱，好像有点

复杂。”

Mr Shu：“其实不复杂，我们常用的概率分布也就那么几种。我们主要记住它们之间的联系并了解正态分布与Gamma分布，连续性分布也就了解得差不多了。”

常用的几种分布如图1-93所示。

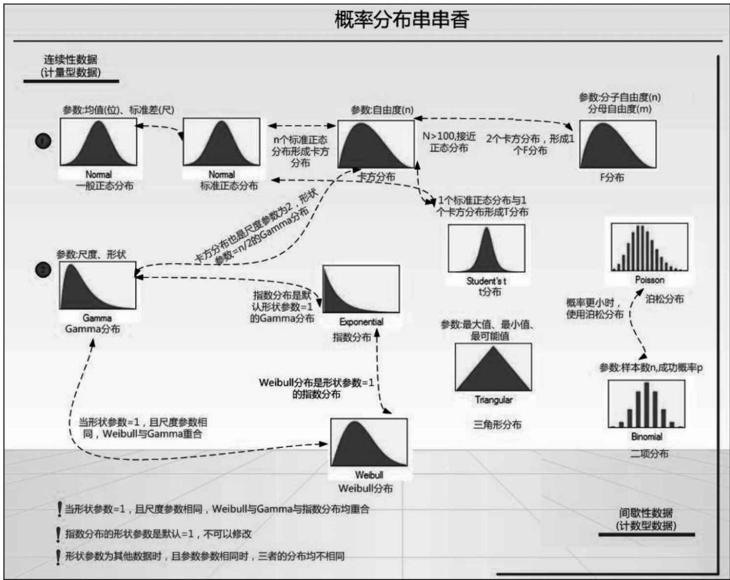


图1-93　常用的几种分布

1.6　做最优秀的面包店长

【场景再现】

Mr shu：“了解概率的一些概念和应用之后有没有觉得概率与预测之间存在千丝万缕的联系？”

Miss Ju：“没错，我觉得生活与工作中其实处处都是概率的应用，要是能列举个实际案例就好了。”

Mr Shu：“在现实生活中经常需要利用预测，如开车出行需要预测到达目的地的时间、父母希望预测子女成年后的身高、农业生产需要预测天气，以及股市投资者需要预测股票的走势等。预测主要是根据过去的情况推断未来的发展，以便及早采取应对措施，争取利益或避免损失。我们就举一个门店经营的案例来说明，在开始之前先问问你小时候有没有过梦想长大以后干什么？”

Miss Ju：“我嘛，最大的理想就是能自己开家店。自己做掌柜，好好经营把生意越做越大，可惜这个理想与我现在的工作太遥远。”

Mr Shu：“其实一点也不遥远，数据预测分析在生活中能处处体现，我们现在就以开店为例子看看我们怎样将数据分析应用在门店运营中？”

Miss Ju：“实在太棒了，这样我的理想和爱好就可以结合在一起了，现在就开始吧。”

Mr Shu：“预测的方法可以分为定性预测与定量预测，定量预测的方法非常多，其中统计分析预测是一种比较常用的预测方法，该方法主要包括回归预测和时间序列预测，前者是截面数据的预测；后者是时序数据的预测。预测的成功与否取决于所用数据是否完备和方法选择是否恰当等，因此任何预测均有风险，来看看预测功能在这些软件中如何实现，首先来看看时间序列预测。”

初识时间序列预测分析

某商场销售人员记录了半年所售商品每天的销售数量，即2013年1月1日～2013年6月30日。他想根据上半年的销售情况预测未来几周或几月该商品的销售情况，操作步骤如下。

（1）整理数据

整理该商品的销售数量数据后，打开Sales工作表，如图1-94所示。

时间	销售量
2013-1-1	1200
2013-1-2	1220
2013-1-3	1227
2013-1-4	1188
2013-1-5	1151
2013-1-6	1190
2013-1-7	1189

图1-94　Sales工作表

（2）选择数据区域

表中数据是某种商品按照每天时间排序的销售量，根据以上数据预测未来几个月的销售数　量。在Excel菜单中单击【Crystal　Ball】选项，单击【Predictor】选项，显示【Predictor】窗口单击【Input Data】选项后设置参数，如图1-95所示。

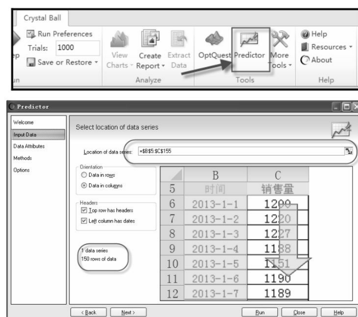


图1-95　设置参数

【Location of date series】下拉列表框自动识别工作表中已存在的数据区域，如果数据不规范，可以手动选择。

【Orientation】选项组中显示数据的方向，即行或列方向。

【Headers】选项组中显示数据表的表头位置，以及选择的数据系列共有多少行数据。确认选择的数据无误后，单击【Next】按钮执行下一步操作。

（3）设置数据属性

操作界面如图1-96所示。

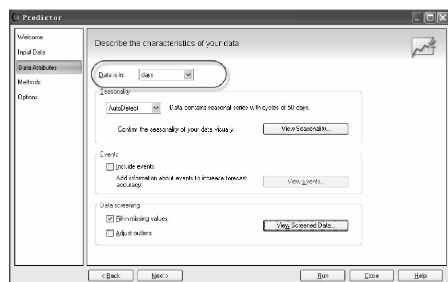


图1-96　操作界面

如果选择的数据有时间周期，如本例中以天为时间单位，则系统会根据日期的差值自动计算出时间单位并显示在【Date is in】下拉列表框中。

【Seasonality】选项组用于选择该数据是否有周期性，如不知道该数据是否有周期性，则单击【View Seasonality】按钮预览数据，如图1-97所示。

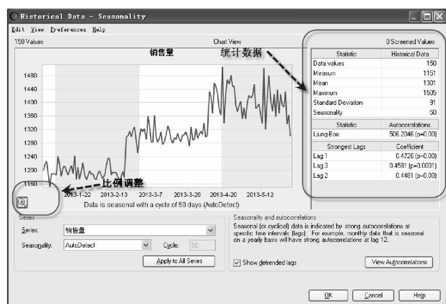


图1-97　预览数据

系统会自动计算该组数据的周期性，本例中的周期性数据为50，即Seasonality＝50，150个数据被分成3个周期。关闭该视图，在【Seasonality】下拉列表框中可以选择【AutoDelect】、【Seasonal】或【Non-seasonal】选项。

本例我们已知道该数据有周期性，所以选择【Seasonal】选项。一般情况下我们可以选择【AutoDelect】选项，让系统自动判断数据的周期性。

如果已知是因为某种原因导致趋势的变化，则可以将原因添入提示事件中。选择【Include Events】复选框，单击【Add】按钮添加说明，如图1-98所示。

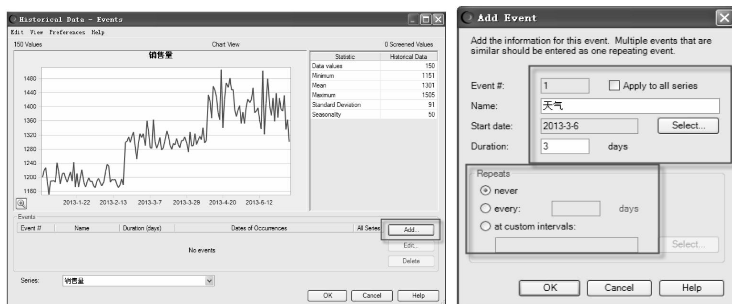


图1-98　添加说明

在本例中3月6日产品销量大增，我们假设是因为反常天气的原因

导致，为此在【Name】文本框中填写事件名称“天气”。

在【Select date】列表框中输入“2013-3-6”，持续3天时间，以后重复的可能性几乎没有。在【Repeats】选项组中选择【Never】单选按钮，系统会自动根据这些信息修正未来的判断。

（4）选择预测方法

单击【Next】按钮，显示【Predictor】窗口，如图1-99所示。

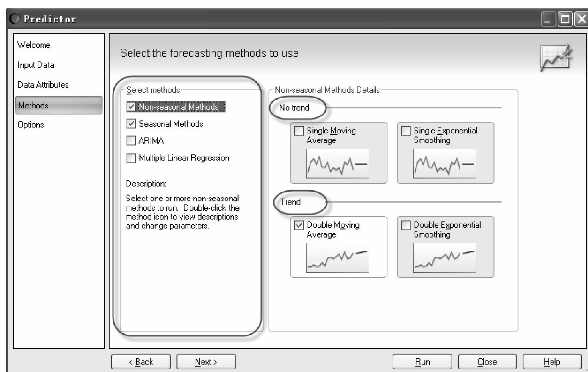


图1-99 【Predictor】窗口

【Non-Seasonal Methods】指非周期性方法，该方法又分为Single Moving Average、Single Exponential Smoothing、Double Moving Average和Double Exponential Smoothing等。

【Seasonal Methods】指周期性方法，该方法又分为Seasonal Additive、Seasonal Multiplicative、Holt-Winters Additive和Holt-Winters Multiplicative等。

Mr Shu："初学者不太好理解这里涉及的预测方法，利用软件的最大好处就是可以不用完全了解这些方法的原则，交由软件自动筛选出最佳的方法。"

Miss Ju："你的意思是全部选择，由软件自动判断哪种方法最适合所选数据，是吗？"

Mr Shu："是的，不过了解一下大致的原理还是可以的；另外，使用时间序列预测还是回归分析预测在Minitab中都可以实现，只是操作不同。"

选择【ARIMA】复选框，系统会自动搜索出最合适的方法，选择后系统提示选择最佳的标准。

选择【Multiple Linear Regression】复选框，当数据有自变量和因变量时选择线性回归方法。

选择【Non-Seasonal Methods】、【Seasonal Methods】和【ARIMA】复选框，单击【Run】按钮开始运算，结果如图1-100所示。

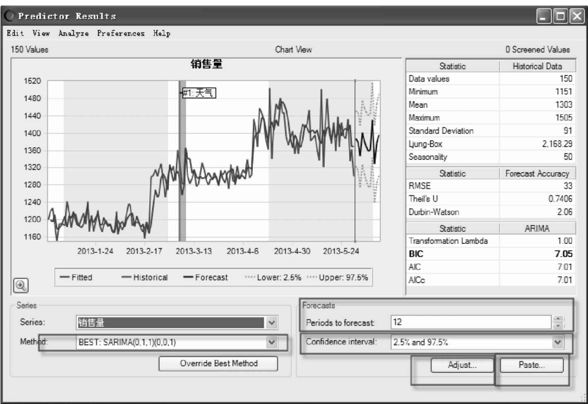
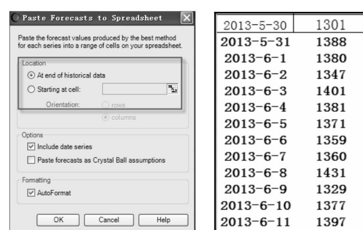


图1-100　运算结果

系统选择“SARIMA(0,1,1)(0,2,1)”预测方法为最佳预测方法，并依此预测，即未来12个周期内在2.5%~97.5%的置信区间内，预测的数据如图中的框出部分所示。

如需要调整数据，则单击【Adjust】按钮。

如需要将预测数据导出到Excel表格中，则单击【Paste】按钮并选择放置的单元格位置，如图1-101所示。



2013-5-30	1301
2013-5-31	1388
2013-6-1	1380
2013-6-2	1347
2013-6-3	1401
2013-6-4	1381
2013-6-5	1371
2013-6-6	1359
2013-6-7	1360
2013-6-8	1431
2013-6-9	1329
2013-6-10	1377
2013-6-11	1397

图1-101　预测数据导出

有关方法说明如下。

（1）Single moving average：一次移动平均，剔除过去几期的平均数据和预测未来值的异常数据，CB　Predictor会自动计算出最佳的若干期平均或者选择若干期的平均数据。这种方法适合用于没有趋势和周期性的易变数据，其结果是直线及平线的预测。

（2）Double moving average：二次移动平均，使用两次移动平均技术，第1次是对原始数据进行移动平均；第2次是对一次移动平均数据的结果再进行一次移动平均。这种方法根据两次移动平均的数据进行评估，CB　Predictor会自动计算出最佳的若干期的平均，或者选择若干期的平均数据。这种方法适用于有一定的趋势性但没有周期性的历史数据，其结果是直线及斜线的预测。

（3）Single exponential smoothing（SES）：单边指数平滑，降低以前数据权重用指数的方法是越近期的数据越重要。这在很大程度上克服了移动平均和百分比变化模型的局限，CB　Predictor可以自动计算平滑指数的最佳情况，或者以手动定义平滑指数。这种方法适用于有一定的趋势性但没有周期性的历史数据，其结果是直线和水平线的预测。

（4）Holt’s double exponential smoothing（DES）：霍尔特双边指数平滑中的双边指数平滑是单边指数平滑两次，第1次对原始数据进行指数平滑；第2次对单边指数平滑的数据再进行指数平滑。CB Predictor在双边指数平滑时使用霍尔特方法，固这种方法在第2次进行指数平滑时与单边指数平滑公式的参数不同。CB　Predictor会自动计算最佳的平滑指数，或者选择自定义平滑指数。这种方法适用于有趋势但是没有季节性的历史数据，其结果是直线和斜线的预测。

（5）Seasonal Additive Smoothing：周期迭加平滑，对没有趋势性的周期性特征的历史数据进行计算，得到预测水平和周期性调整预测值的指数平滑值。这种周期性调整是对预测水平的增强，得到周期迭加预测。该方法适用于没有趋势性但是周期性不随时间增加而增加的历史数据。其结果是曲线预测，表示数据的周期性变化。

（6）Seasonal Multiplicative Smoothing：周期乘积平滑，对没有趋势性的周期性特征的历史数据进行计算，得到预测水平和周期性调整预测值的指数平滑值。这种周期性调整是对预测水平的增强，得到周期乘积预测。该方法适用于没有趋势性但是周期性不随时间增加而增加的历史数据。其结果是曲线预测，表示数据的周期性变化。

（7）Holt-Winters’Additive Seasonal Smoothing：霍尔特周期迭加平滑，是对有周期特征的霍尔特指数平滑的扩展。这种方法在3个

方程式基础上建立，得到预测水平、预测趋势和周期性调整预测的指数平滑值。这种周期迭加方法增加了趋势预测周期性因素，得到霍尔特迭加预测。该方法适用于有趋势性并且周期性不会随着时间增加而增加的数据，其结果是曲线预测，表示数据的周期性变化。

（8）Holt-Winters’ Multiplicative Seasonal Smoothing：霍尔特周期乘积平滑，这种方法与霍尔特周期迭加方法相似，对预测水平、趋势和周期性调整计算出指数平滑值。这种周期乘积方法对预测周期性因素进行相乘，得到霍尔特周期乘积平滑。该方法适用于有趋势性并且周期性会随时间增加而增加的数据，其结果是曲线预测，表示数据的周期性变化。

初识回归预测分析

在预测分析中还有一类使用较为广泛的方法是回归预测，因为应变量影响多个自变量，所以预测的结果需要回归。这里以北方居民使用采暖量为例说明回归预测分析。

供暖公司为了预测某地区的取暖量，以月份为单位收集了从2006年～2010年的供暖量及影响采暖量的3个因素的数据。该公司认为影响该地区供暖量主要有3个因素，即新房数量、平均气温及暖气的单价，有关数据如图1-102所示。

日期	用热量	新房数量	平均气温	暖气单价
2006-1-1	92.00	151	0	￥6.4
2006-2-1	53.00	128	-1	￥6.2
2006-3-1	84.00	85	5	￥6.0
2006-4-1	54.00	52	7	￥6.3
2006-5-1	5.00	5	19	￥5.5
2006-6-1	63.00	134	21	￥5.2
2006-7-1	46.00	92	22	￥6.2
2006-8-1	40.00	171	24	￥6.8
2006-9-1	72.00	248	18	￥7.0
2006-10-1	59.00	212	11	￥7.4
2006-11-1	104.00	268	4	￥7.4
2006-12-1	78.00	226	5	￥7.5
2007-1-1	119.00	146	2	￥7.7
2007-2-1	57.00	124	2	￥8.3
2007-3-1	49.00	97	4	￥7.1
2007-4-1	149.00	403	10	￥7.0
2007-5-1	28.00	285	18	￥6.6
2007-6-1	56.00	315	23	￥6.2
2007-7-1	62.00	346	22	￥7.3
2007-8-1	57.00	247	21	￥7.8
2007-9-1	48.00	288	17	￥7.6

图1-102有关数据

使用回归预测分析的操作步骤如下。

（1）选择数据

打开“采暖回归分析.xls”工作表，在Excel菜单栏中选择【Crystal Ball】选项。单击【Predictor】选项，系统自动选择数据，如图1-103所示。

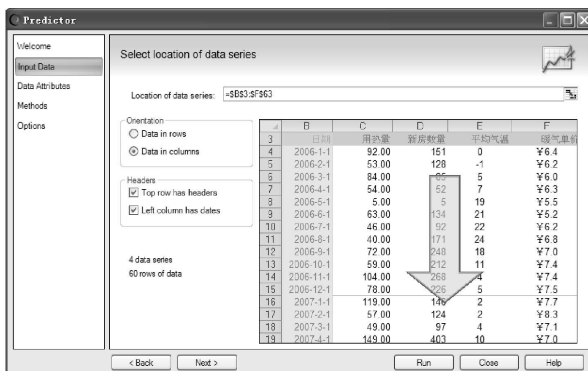


图1-103　选择数据

单击【Next】按钮，系统自动设置数据属性。

（2）选择回归预测方法

选择回归预测方法，如图1-104所示。

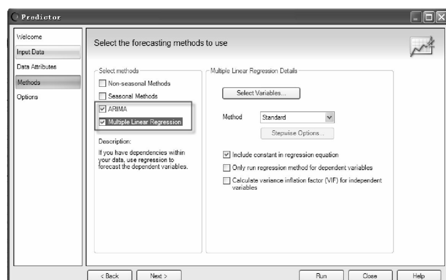


图1-104　选择预测方法

必须选择除回归分析之外的其他分析方法，因此选择【ARIMA】和【Multiple Linear Regression】复选框。单击【Next】按钮，弹出【Regression Variables】对话框，如图1-105所示。

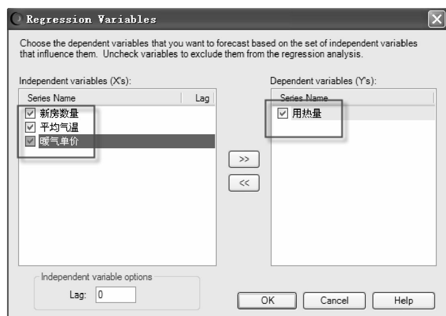


图1-105　【Regression Variables】对话框

在【Independent variables】选项组中选择3个自变量，即“新房数量”、“平均气温”和“暖气单价”；在【Dependent variables】选项组中选择因变量“用热量”。

（3）查看预测结果

单击【OK】按钮，分析结果如图1-106所示。

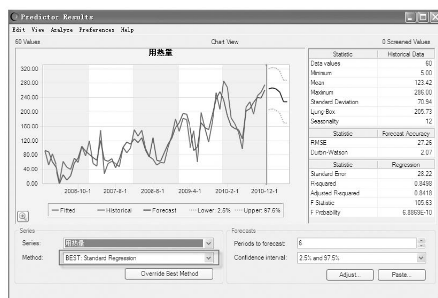


图1-106　回归预测分析结果

在【Method】下拉列表框中系统提示最佳的预测方法为“Standard Regression”，在【Series】下拉列表框中可以选择其他3个变量的预测结果及预测方法，这些变量的最佳预测方法各不相同。

（4）生成分析报告

在分析出结果后与时间序列预测一样可以将分析结果导入Excel表格中，并且自动生成分析报告，在图1-106所示窗口中选择【Analyze】|【Create Report】选项，显示【Create Report Preferences-Predictor】对话框。选择【Predictor】选项，如图1-107所示。



图1-107　选择【Predictor】选项

单击【OK】按钮，生成报告，其中的一部分如图1-108所示。

Crystal Ball Report - Predictor
Created 2014-12-1 at 16:36

Summary:	
Data attributes:	
Number of series	4
Data is in	months
Run preferences:	
Periods to forecast	6
Fill-in missing values	On
Adjust outliers	Off
Methods used	Multiple Linear Regression
	ARIMA methods
Forecasting technique	Standard forecasting
Regression technique	Standard
Constant included	On
Error measure	RMSE

图1-108　部分报告

Mr　Shu：“我们已经了解了Crystal　Ball预测的基本功能，并且知道了使用时间序列和回归分析来预测。”

Miss　Ju：“哇哦，真的是好方便。还可以自动生成报告，省去一些写报告的时间。”

Mr Shu：“是的，前面我们介绍了时间预测与回归分析的基本操作。你不是说你的梦想是拥有一个属于自己的店吗？现在就让你来体验一把做店长的瘾。其实经营一家门店也需要学问，如通过优化的思维分析门店的经营。下面我们就将通过你的面包店来回顾前面讨论的知识点，你想好店名了吗？”

Miss Ju：“名字不重要，我主要想看看怎么当好店长？”

Mr Shu：“那就开始吧。”

做最优秀的面包店长

花小姐的面包店是一家位于上海浦东区且迅速增长的面包店，它设立于2007年3月。花小姐是一个非常细心的店长，从开业以来一直在Excel工作簿中仔细记录店内3种主要产品的销售数据，即法式面

包、意大利式面包和匹萨。经过几年的经营积累，她的门店已经小有规模。现在她想改进，但是受库存地点限制必须预测未来的产品市场，并依此对人员和库存等进行战略性和长远的决策。决策的依据基于她所做的数据积累，即通过分析数据中的规律来改进。

花小姐预测的最初目的是要保持足够的原料，以满足店内生产的要求。以往面包原材料会 定期向供应商购买，并在大量购买时得到折扣。如果店内产品销售过旺，原材料就会紧缺；反之会有多余库存。所以必须保持库存和产品的平衡，以保证产品始终用最新鲜的配料来进行生产。

3种产品需要的原料大致一样，主要是面粉、酵母和食盐。如果不预测市场，就会导致原材料的需求量忽高忽低。原材料供应商也有可能会因此提高价格，所以预测产品市场不仅仅能保证材料的新鲜度，还能最大程度地降低成本。

有了对产品市场的预测，花小姐需要购买原材料时也能保证其产品的质量，因此需要有效地预测未来的销售收入。她在Excel电子表中记录了每种产品从2007年3月份开始至今的日常销售数据并保存在“面包店经营”工作簿的“销售数据”工作表中，如图1-109所示。

日期	法式面包	意大利面包	披萨	总量
1-Mar-07	$591.64	$852.96	$187.04	$1,631.64
2-Mar-07	$582.36	$1,101.63	$187.29	$1,871.28
3-Mar-07	$591.63	$1,008.88	$186.79	$1,787.31
4-Mar-07	$589.60	$1,071.82	$186.72	$1,848.14
5-Mar-07	$582.48	$1,038.78	$186.88	$1,808.14
6-Mar-07	$579.34	$1,002.81	$186.95	$1,769.10
7-Mar-07	$606.50	$995.83	$187.13	$1,789.46
8-Mar-07	$611.57	$837.60	$187.07	$1,636.25
9-Mar-07	$605.11	$1,125.16	$187.29	$1,917.55
10-Mar-07	$617.31	$1,053.87	$186.93	$1,858.11
11-Mar-07	$630.87	$1,143.00	$187.50	$1,961.37
12-Mar-07	$656.43	$1,049.61	$187.04	$1,893.08
13-Mar-07	$667.51	$1,271.85	$187.42	$2,126.78
14-Mar-07	$658.08	$1,140.49	$187.08	$1,985.65
15-Mar-07	$670.71	$1,174.30	$187.29	$2,032.30

图1-109 “销售数据”工作表

花小姐以表中的原始数据为基础，将自2007年以来的原始数据整理为3种产品以周为时间周期的数据。周产品销售数据保存在“运营”工

作表中，并且注明了原料的名称。通过创建这个数据表花小姐想对未来几周的产品的销售情况进行预测，周销售数据表如图1-110所示。

	数据		
周数	法式面包	意大利式面包	披萨
1	4,123.55	7,072.71	1,308.81
2	4,446.88	7,621.58	1,310.32
3	4,663.57	7,899.54	1,311.78
4	5,013.63	7,404.53	1,312.39
5	5,451.94	6,963.79	1,312.79
6	5,536.90	6,658.48	1,314.58
7	5,673.72	7,270.59	1,314.30
8	5,813.07	6,558.68	1,314.69
9	5,922.39	6,736.44	1,315.31
10	6,214.79	6,514.98	1,316.65
11	6,495.31	7,057.55	1,317.85
12	6,582.71	6,904.19	1,318.61
13	6,848.72	8,422.65	1,320.23
14	6,863.34	8,314.42	1,322.54
15	6,904.93	7,399.92	1,324.36
16	6,939.59	7,370.00	1,324.82
17	7,291.25	6,681.32	1,325.14
18	7,391.09	6,152.72	1,327.17
19	7,296.55	6,400.72	1,327.37

图1-110　周销售数据表

该面包店已经收到这个月的订货，花小姐必须要在这个月确定本月和下个月的原材料订单，因此必须预测未来两个月内的销售。她现在有173周的销售数据，需要预测未来8周的销售数据。

（1）建立Excel模型

在未来两个月花小姐没有调整产品价格的计划，每种产品的单位质量和单价不变，因此预测原料的需求量首先要知道3种商品的销售量。建立该数学模型的思路为：商品销售预测→商品重量预测→原材料预测，在Excel建立的数学模型如图1-111所示。

174 ← 预测开始周

4周预测

要求	销售收入预测	销售价格	单位	单位重量	总重量
法式面包	$ -	$ 1.24	-	0.25	-
意大利式面包	$ -	$ 2.08	-	1.05	-
披萨	$ -	$ 14.71	-	3.12	-

原料	预测重量	每磅	磅数要求
法式面包	0		
面粉		0.32	0
酵母		0.0010	0
食盐		0.0015	0
意大利式面包	0		
面粉		0.36	0
酵母		0.0015	0
食盐		0.0010	0
披萨	0		
面粉		0.48	0
酵母		0.01	0
食盐		0.00	0
西红柿		0.10	0

图1-111　建立的数学模型

说明如下。

·单元格B39：E213区域为2007年3月份以来3种产品每周的销售数据。

·C9单元格用于统计预测的未来4周内法式面包的销售收入，在其中输入“＝SUM(INDEX($B$41:$E$299,$C$3,2):INDEX($B$41:$E$299,$C$3＋3,2))”。

·在C3单元格内输入开始的周数，初始设置为174，即最后一周。

·C10单元格用于统计预测未来4周内意大利式面包的销售收入，C11单元格用于统计预测未来4周内匹萨的销售收入。

·D9:D11单元格区域内为每种商品的销售单价，这样用销售收入除以单价即可知道销售数量。

·在E9单元格内输入公式“＝C9/D9”，其他依此类推；F9:F11单元格为每种商品的单位重量，数量乘以单位重量可以知道每种商品的重量；在G9单元格内输入公式“＝E9*F9”，其他依此类推。

·B14:E27单元格区域计算每种商品需要的原料，按照每种商品需

要的原料组成计算；在C15单元格内引用G9单元格数据；在E16单元格内输入公式“＝$C$15*D16”计算法式面包需要的原料面粉的数量，其他原料成分计算依此类推；在D31单元格内输入公式“＝SUM(E16,E20,E24)”将3种商品的面粉原料求和，这是需要供应商提供的原料采购的数据。

（2）预测设置

选择B39:E213单元格区域内的任一单元格，选择Crystal Ball菜单中的【Predictor】选项，显示的【Predictor】选项如图1-112所示。

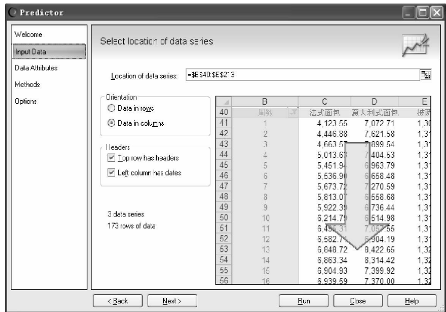


图1-112　【Predictor】选项

系统自动选择数据表格所在的位置，单击【Next】按钮，选择【Data Attributes】选项，如图1-113所示。

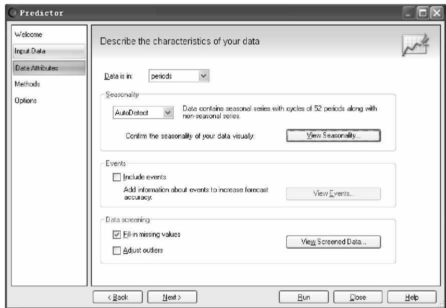


图1-113　【Data Attributes】选项

保留系统默认值，单击【Next】按钮，显示【Methods】视图，如图1-114所示。

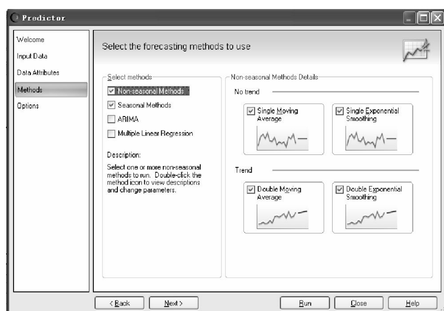


图1-114　【Methods】视图

该视图主要用于设置数据预测的方法，有时间序列的数据选择【Non-seasonal Methods】和【Seasonal Methods】选项。

（3）查看分析结果

在【View】下拉菜单中选择有关选项查看各产品的销售情况，法式面包明显有趋势而无周期；意大利式面包既有周期，也有趋势性。为了预测准确，选择所有预测方法，由系统来确定最佳的方案。选择【ARIMA】复选框，单击【Run】按钮，结果如图1-115所示。

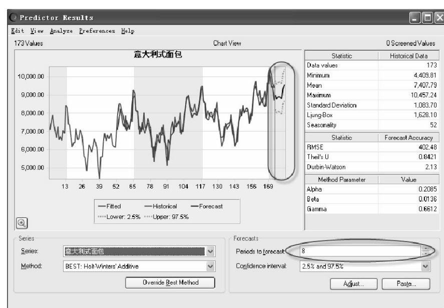


图1-115　预测结果

每种产品的预测数据不同，使用的方法也不同。在预测周期【Periods to forecast】微调框中 设置8，即预测8个周期。【Method】下拉列表框中显示最佳的分析方法，单击【Paste】按钮保存预测结果，如图1-116所示。

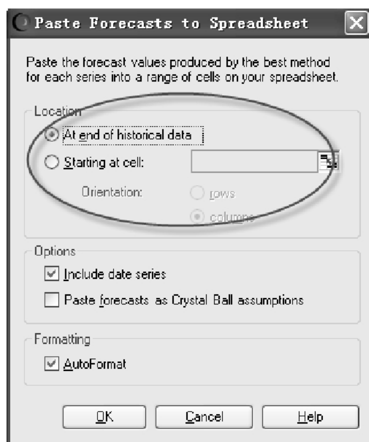


图1-116　保存预测结果

在【Location】选项组中选择将预测数据放在原历史数据的后面或指定单元格区域，选择【At end of historical data】单选按钮。单击【OK】按钮，3组预测数据复制到“运营”工作表中的数据表中，如图1-117所示。

	数据		
周数	法式面包	意大利式面包	披萨
173	18,596.88	9,371.57	1,483.24
174	18,674.20	8,991.94	1,476.13
175	18,751.40	8,782.64	1,469.01
176	18,828.61	8,872.59	1,461.88
177	18,905.81	8,898.17	1,454.76
178	18,983.02	8,829.41	1,447.64
179	19,060.23	8,794.43	1,440.52
180	19,137.43	9,320.40	1,433.40
181	19,214.64	9,550.62	1,426.28

图1-117　预测后的数据表

3种商品的预测重量及原料的采购数量在数据模型中均已完成计

算，如图1-118所示。

要求	销售收入预测	销售价格	单位	单位重量	总重量
法式面包	$ 75,160	$ 1.24	60,612.92	0.25	14,850.17
意大利式面包	$ 35,545	$ 2.08	17,089.11	1.05	17,943.57
披萨	$ 5,862	$ 14.71	398.49	3.12	1,243.29

统一要求	磅数
面粉	11,809
酵母	54
食盐	41
西红柿	124

图1-118　原材料采购数量

根据在模型中预测计算出未来两个月的原材料需求量，此时一定会根据现有的库存和原材料的新鲜程度来指定最佳的订货数量。

现金流对于门店经营的重要性不言而喻，花小姐也会详细记录每个月的现金流。这样不仅 可以帮助她管理预估库存，并且用它来预测门店的收入使她的现金流动情况变得更好，更好地了解面包店的现金流量会帮助其更好地控制主要资本支出。如果花小姐想在门店内新增设备或仓库等，则必须要了解接下来几个月的现金流情况。

简单来讲，现金流就是除去开支每月的剩余资金。如果用公式来解释，就是销售收入-门店成本和其他开支。门店成本主要包括商品成本和税赋成本，商品成本中又包括固定成本和变动成本。这需要我们建立数据模型，其他开支是花小姐扩大规模带来的那些支出。

花小姐认为主要有两个方面的支出，即面粉和运费。她想开始在7月份囤积一些油，为此需要增加一个筒仓。并且在8月份购买一辆新的面包车以方便在附近社区送货，她需要预测何时可以实施这些项目计划或是否需要再等一段时间。

在“现金流”工作表中给出了面包店从2007年以来的现金流量情况，并且花小姐将3种主要商品的销售数据按照月份为周期制作了一个数据透视表。当然以月份为周期的销售数据也是基于日销售表的基础上计算出来的，可见原始数据的积累是多么重要。现在她需要预测未来3个月的收入来计算现金流的情况后决定费用的支出，并且为了保证门店的正常运营，每月末店内的净现值必须大于20 000美元。

操作步骤如下。

（1）建立Excel数据模型

确定现金流首先要确定各成本，成本由商品与税费成本组成。每类成本又由固定成本与可变成本组成，两类成本的固定成本均已知。只有变动成本不知，而它均与销售收入有关，因此该数学模型的思路为收入预测→计算成本→每月现金流→决策。

在Excel中的“现金流”工作表中建立模型，如图1-119所示。

	Common-Sized	7月	8月	9月
收入预测	100%	$ -	$ -	$ -
费用	**Common-Sized**	**7月**	**8月**	**9月**
商品成本				
固定成本	$ 6,707.60	$ 6,707.60	$ 6,707.60	$ 6,707.60
可变成本	23%	$ -	$ -	$ -
Overhead				
固定成本	$ 8,924.00	$ 8,924.00	$ 8,924.00	$ 8,924.00
可变成本	18%	$ -	$ -	$ -
税收	5%	$ -	$ -	$ -
增值税	17%	$ -	$ -	$ -
总费用		$ 15,631.60	$ 15,631.60	$ 15,631.60

图1-119　建立模型

现金流的Excel模型说明如下。

单元格B33:AP36区域为2007年3月开始以月度为时间周期的历史销售收入数据。

E4:G4单元格区域为预测未来3个月的销售收入数据。

B8:G16单元格区域为每个月店内的成本。

成本包括商品成本和间接成本，商品成本主要指原料的采购成本。其中的固定成本指店面租金等，为$6707/月。商品可变成本与销售收入有关，按照经验估计可变成本占销售收入的23%。在E10单元格内输入公式“＝$D10*E$4”，即7月份的商品可变成本。其他月份商

品的可变成本依次类推；间接成本主要包括设备折旧等费用，为$8924/月。按照经验间接可变成本占销售收入的比例约为18%。税收比例为5%，增值税比例为17%。

在E13单元格内输入公式“＝$D10*E$4”表示7月间接可变成本费用。

在E14单元格内输入公式“＝E$4*$D14”表示7月份的税收费用。

在E15单元格内输入公式“＝E$4*$D15”表示7月份增值税的费用。

在E16单元格内输入公式“＝SUM(E8：E15)”表示7月份店内的总费用。

其他月份的间接成本计算依此类推。

7月份计划囤油需要筒仓，需投资$50 000，数据输入至E20单元格；8月份新购面包车及新增仓库施工的一次性投资为$35 000，数据输入至F21单元格。每月的现金流＝销售收入-总费用-投资。在E24单元格内输入公式“＝E4-E16-SUM(E20:E21)”表示7月份的现金流。假设7月初的净现值为$42 941，则输入至E26单元格。在E27单元格内输入公式“＝E26＋E24”表示7月末的净现值，其他月份依此类推。

（2）预测设置

由于现金流的预测依然按照时间序列分析方法进行，因此在Crystal Ball中设置预测器的方法与上面案例相同。操作步骤与库存控制相同，如图1-120所示。

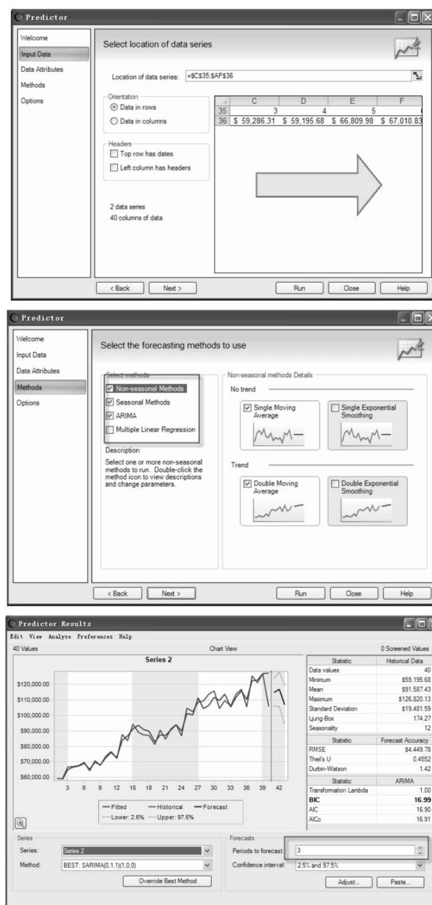


图1-120　设置预测器

此时预测周期为3，即只需要预测未来3个月的销售收入。预测完成后将预测数据放置在表格最后，如图1-121所示。

7	8	9
$114,963.84	$116,873.30	$106,861.40

图1-121　预测结果

预测完成未来3个月的销售收入。按照Excel的数学模型，如果7月份需要投资$50　000，8月份需要投资$35　000且7月份的月初净现值$42 941计算，则每月月末的净现值如图1-122所示。

Extraordinary Items	7月	8月	9月
筒仓投资	$ 50,000.00	$ -	$ -
新的面包车及施工	$ -	$ 35,000.00	$ -
每月现金流	$ (23,404.66)	$ (7,703.30)	$ 23,619.26
月初的净现值	$ 42,941.00	$ 19,536.34	$ 11,833.04
月末的净现值	$ 19,536.34	$ 11,833.04	$ 35,452.31
最小的现金目标	$ 20,000.00	$ 20,000.00	$ 20,000.00

图1-122　每月月末的净现值

从计算结果来看，9月末的净现值$35　452满足最低现金目标$20 000的需求。但8月末的净现值$11 833不能满足最小现金目标，7月末的净现值$19　536也与最小现金目标接近。这些数据均是Excel中单个数据的计算结果，不能代表现金流的风险。门店管理者要知道的是风险的概率、因此需要设置假设变量。

（3）设置假设变量

在现金流中的主要不确定因素有商品成本中的可变成本的比率、间接成本中的可变成本的比率及税收的比率；另外，还有一个重要的不确定因素是预测的销售收入。该输入也是一个数据概率，而不仅仅是一个数值，因此我们需要设置以上假设变量。在Crystal　Ball预测结束后可以直接将预测结果设置为假设变量，并使用时间序列分析的预测值序列。CB　Predictor默认会得到一个正态分布的假设，假设变量的设置如图1-123所示。

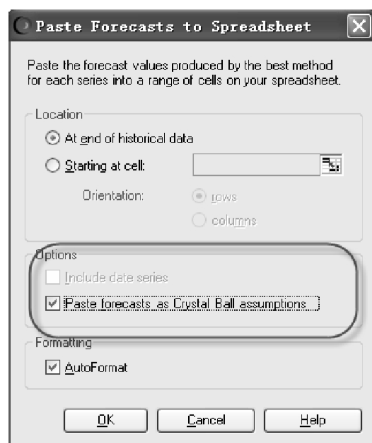


图1-123　设置假设变量

在预测运行之后单击【Paste】按钮粘贴数据时选择【Paste Forecasts as Crystal Ball assumptions】复选框，预测值自动设置成以单元格数据为均值的正态分布，如图1-124所示。

7	8	9
$114,963.84	$116,873.30	$106,861.40

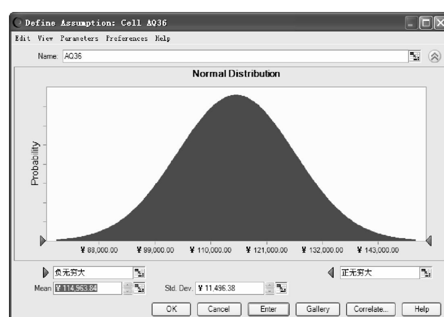


图1-124　预测值自动设置成以单元格数据为均值的正态分布

将商品成本中的可变成本、税赋中的可变成本及增值税率设置为假设变量，如图1-125所示。

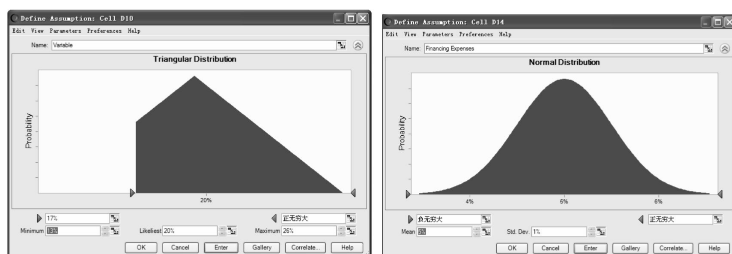


图1-125　设置其他假设变量

D10单元格设置最小值为13%，最大值为26%，最可能值为20%的三角形分布；D13单元格设置成均值为12%，标准差为1%的正态分布；D14单元格设置成均值为5%，标准差为1%　的正态分布；D14单元格设置成均值为22%，标准差为2%的正态分布。

（4）设置目标函数

花小姐最终想知道的是每月末的净现值大于最低风险值$20　000的概率值是多少，而不仅仅需要其中一个数据。此时需要将每月末的净现值设置为目标函数，即预测值。操作过程略，设置结果如图1-126所示。

Extraordinary Items	7月	8月	9月
筒仓施工	$ 50,000.00	$ -	$ -
新的仓库	$ -	$ 35,000.00	$ -
每月现金流	$ (22,875.49)	$ (7,875.49)	$ 27,124.51
月初的净现值	$ 42,941.00	$ 20,065.51	$ 12,190.03
月末的净现值	$ 20,065.51	$ 12,190.03	$ 39,314.54
最小的现金目标	$ 20,000.00	$ 20,000.00	$ 20,000.00

图1-126　设置目标函数的结果

将E27单元格设置为名称为“7月净现值”的预测变量，将F27单元格设置为名称为“8月净现值”的预测变量，将G27单元格设置为名称为“9月净现值”的预测变量。

这样较为完整的数学模型即设置完成了，可以进行风险模拟分析。

（5）风险模拟分析

单击【Start】按钮，分析结果如图1-127所示。

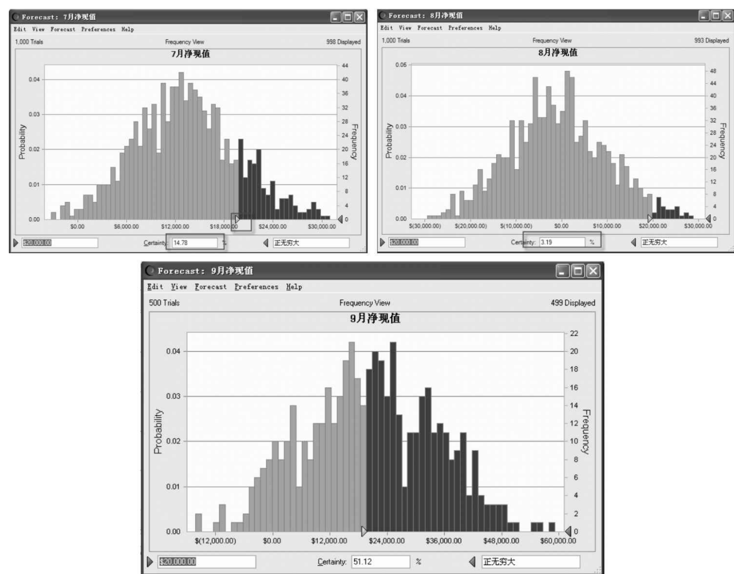


图1-127　风险分析结果

查看每月末净现值大于$20　000的概率，在每个月份的净现值概率分布图中左侧输入“20 000”后按回车键。

从以上分析可以看出，7月和8月的现金流风险较高。特别是8月份高于$20　000的概率非常低，为3.17%；7月的风险也不低，高于安全线的概率为15%。

花小姐可以清晰地知道8月份的风险最高，可以考虑将计划延迟。即将7月份筒仓的投资分批为两个月进行；8月份购置新车的计划也可以延期到9月份进行，这就是定量风险分析给我们的结果。

花小姐的面包店是一种劳动密集型的运作，需要大量的员工，目前已有18名。然而为了维持目标利润，她必须控制自己的劳动力成

本。她需要购买很多包括揉、捏及混合面等的昂贵机器，因此她更需要准确地预测其劳动力成本。这样可以决定何时投资新增设备，以保持在预算之内她的总开支可控。

花小姐知道在所处的地区一些关键的宏观经济数字对劳动力成本有影响，如工业生产指数、当地物价指数和当地的失业劳动力成本，这些数字在互联网上每月都由统计局和商务部发布。

花小姐对人力成本做了一个数据图表，其中列出每个月平均每小时的工资和每个月其他3个因素的数值。

现在来分析未来几个月内的劳动力成本是否还有上升的空间，花小姐是否能新增一些机器来替代人工。操作步骤如下。

（1）建立劳动力成本模型

花小姐在Excel中罗列了人员工资和其他成本的信息，如图1-128所示。

面包店内劳动力成本分析

	Jun-10	Dec-10
工资	$ 14.27	$ -
其他	33%	33%
全部雇员	18	18
全部雇员成本	$ 54,652.74	$ -
雇员成本变化		-100%

从回归分析的经济指标

Date	平均工资	工业生产指数	本地CPI	本地失业率
Jun-07	$13.22	121.9	469.0	8.0
Jul-07	$13.19	122.6	469.3	8.4

图1-128　劳动力成本模型

相对于库存模型和现金流模型，劳动力成本模型相对简单，设置说明如下。

雇员成本＝人数×小时工资×(1＋税率)×小时数，因此在D7单

元格中输入“＝D4*20*8*(1＋D5)*D6”，表示6月份的全部雇员的人力成本；同样E7单元格表示10月份的人力成本，而10月份的小时工资未知，下面我们需要对10月份的小时工资进行预测分析。

（2）预测未来小时工资

员工每个小时的工资由其他3个数据决定，因此花小姐决定使用回归分析的方法，而不是时间序列的方法。对于回归分析，因变量是花小姐的平均工资，其他3个因素是自变量。

操作的前几步与时间序列分析相同，这里不再叙述。

选择回归分析方法，如图1-129所示。

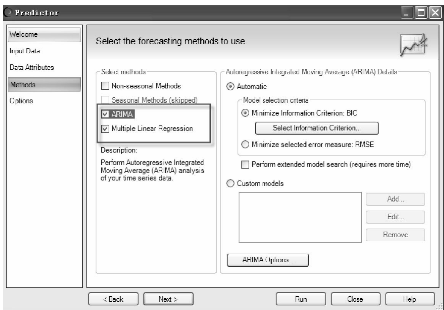


图1-129　选择回归分析方法

选择【Multiple Linear Regression】复选框，单击【Run】按钮，预测结果如图1-130所示。

Jul-10	$ 14.59	126.8	500.7	7.6
Aug-10	$ 14.51	125.9	500.8	7.7
Sep-10	$ 14.45	125.2	500.8	7.9
Oct-10	$ 14.40	124.7	500.8	8.0
Nov-10	$ 14.36	124.3	500.8	8.1
Dec-10	$ 14.33	124.1	500.8	8.2
Coefficients for ·	0.0000	0.0571	0.0175	-0.1846

图1-130　平均工资预测结果

（3）查看分析结果

预测小时平均工资后我们可以查看人力成本的变化率，如图1-131所示。

面包店内劳动力成本分析				
		Jun-10		Dec-10
工资	$	14.27	$	14.33
其他		33%		33%
全部雇员		18		18
全部雇员成本	$	54,652.74	$	54,875.86
雇员成本变化				0.41%

图1-131　人力成本分析结果

从以上分析可以看出在未来6个月期间小时单位工资由$14.27上升到$14.33，增长幅度仅为0.41%。通过这些结果花小姐可以认为人力成本不会增加，未来6个月内有足够的资金购买新的设备。

当然这只是其中一个预测结果，可以与现金流分析一样将预测结果设置为假设变量后进行模拟分析。

Miss　Ju：“哇，想不到当个面包店长还有这么多的学问。需要积累原始数据、建立数学模型和预测分析等，真是大开眼界。”

Mr Shu：“那当然，可不要小看开店，凡事皆有学问。其实管理企业与管理门店大同小异，只要思路对了，效益自然会产生。我们选择还只是在数据分析上做了一小步，还有更多更好的应用等着我们。”

Miss　Ju：“真的？但是我都觉得有点头脑发晕了，要不要先整理一下。”

Mr Shu：“需要，只有不断总结并消化吸收我们才可以触类旁通，现在总结第1章的内容如下。”

我们首先从不靠谱的世界杯开始接触概率分布的概念，通过手机套餐的案例了解了概率分布在定量风险分析的应用；通过笔记本电脑出国冒险记案例对定量风险分析进一步加深了印象；从标准正态分布开始拓展接触了其他一些常用的概率分布，并对分布之间的关联进行了分析。

了解概率后我们通过花小姐的面包店的案例讨论了预测，包括时间序列预测和回归分析预测。

本章在风险定量分析和预测分析的应用上主要通过Crystal Ball操作，在模拟分布时主要通过Minitab操作。

Crystal Ball除了应用在预测和定量风险分析外，优化分析也是其强项之一。在第2章中我们主要通过解释线性规划的原理和软件的操作来理解如何进行优化分析；同时应用在优化计算上的软件也有很多，第2章将主要介绍Excel、LINGO和Crystal Ball等，说明它们在数据分析中的应用。

第2章
运筹帷幄，决胜千里
效益最大化

2.1　换个思路来数鸡

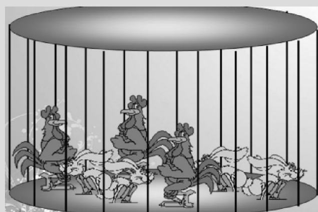
2.2　做一个精明的农场主

2.3　见识LINGO与Crystal Ball的威力

2.1　换个思路来数鸡

［古］今有雉兔同笼，上有十五头，下有四十足，问雉兔各几何？

［今］现在有一群鸡兔同笼，共有15个头和40只脚，问鸡和兔各有几只？



Miss Ju："Mr Shu，你这是想考我小学的算术题吗？有多少只鸡和兔还不会算吗？"

Mr Shu："做这道题对你来说肯定是大材小用，但是我的目的不是让你真正解答这个问题，而是想通过这道题来引申出我们第2章的主要内容——优化求解。"

Miss Ju："优化求解与这个例子有什么关系？直接用方程求解不就可以了吗？难道这题还有其他方式可以更方便求解吗？"

Mr Shu："当然有，但是我们的重点也不是在求解这道题，而是通过这道题来认识一些优化求解的工具，如Excel、Crystal Ball、Lingo和Minitab等。"

Miss Ju："Excel还有优化求解的功能？Lingo又是什么软件？"

Mr Shu："我们在本章重点介绍这些软件在优化计算方面的一些应用，本来是应该先了解方差分析及回归分析等常规方法之后再来应用优化软件计算。不过没有绝对的前后关系，我们先看看优化计算在工作和生活上是如何应用的。"

Miss Ju："听起来好像不错，那我们现在直奔主题吧。"

Mr Shu："不要着急，我们先换个思路来求解鸡兔的问题。"

【鸡兔同笼新解】

假设鸡和兔子训练有素，能听明白人的指令，现在我们吹一声口哨代表要求它们抬起一只脚。鸡兔总数有15只，各抬起一只脚，剩余的脚数为40-15＝25只。再吹一声哨子代表再抬起一只脚，剩余的脚数为25-15＝10只。兔子抬起两只脚没有问题，而鸡就一屁股坐地上了。那剩下的10只脚为兔子的，兔子的个数为10/2＝5只，鸡的个数就为15-5＝10只。

Miss Ju：“这样也可以解决问题，真的是很神奇。”

Mr Shu：“是的，有没有难度，关键是思路。当然这些问题都比较简单，可以通过一些比较有趣的方法来求解。但问题一旦复杂，就没那么好解了。”

下面我们看看如何使用软件来求解这些问题。

案例：**Excel**规划求解——鸡兔同笼

Excel是最常用的数据整理和分析的工具，在数据分析方面提供了模拟运算表、方案管理器、单变量求解、规划求解和数据透视表等工具。我们先了解其最神奇的一项功能，即规划求解。

线性规划是在一定的限制条件下利用数学的方法运算使前景的规划达到最优的方法。它是现代管理科学的一个重要手段，也是运筹学的一个分支。

线性规划的运算方法很多，如图解法、表上作业法和图上作业法等。其中最常用的是单纯形法，Excel的规划求解功能就是按照它设计的，该方法是应用数学迭代的原理求解线性规划的一种基本方法。线

性规划的应用范围非常广泛，包括工农业生产、交通运输、财务金融及决策分析等。使用Excel的规划求解工具，即使不理解其数学原理和公式推导，只需要学会操作也可以轻松求解结果。

规划求解需要Excel的加载宏，加载该宏之后可以利用Excel的规划求解功能进行规划求解。加载规划求解宏的方法如下。

（1）在Excel　2003版本中单击【工具】|【宏】|【加载宏】选项，加载【规划求解加载项】即可加载该宏。

（2）在Excel　2007版本中单击【Office】按钮，【Excel选项】|【加载项】|【转到Excel加载项】选项，然后加载【规划求解加载项】即可加载规划求解的宏。

（3）在Excel　2010版本中单击【文件】选项，打开【Excel选项】对话框。单击【加载项】 标签，单击【转到】按钮，选择【规划求解加载项】复选框，单击【确定】按钮即可在工具栏的【数据】选项卡中出现【分析】选项组，其中就有【规划求解】按钮，如图2-1所示。

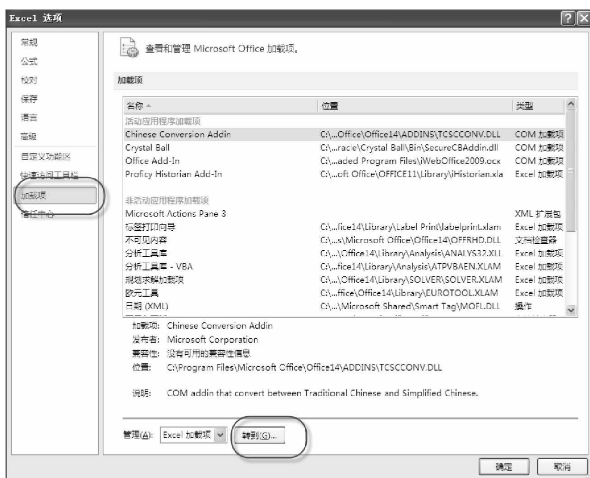


图2-1　【加载项】标签的【转到】按钮

显示【加载宏】对话框，如图2-2所示。

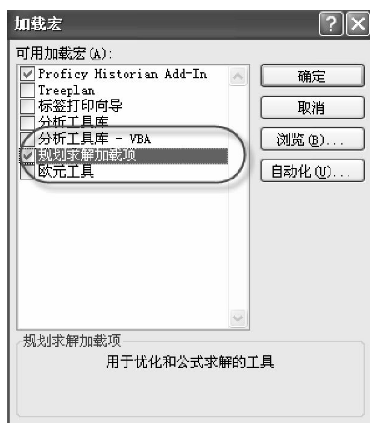


图2-2 【加载宏】对话框

选择【规划求解加载项】复选框，单击【确定】按钮即可在工具栏的【数据】选项卡中出现【分析】选项组，其中就有【规划求解】按钮。

规划求解基本上能涵盖模拟计算器、单变量求解和方案管理器等工具的功能，并且能十分方便快捷地计算。

将已知条件修改为共35只鸡和兔，94只脚，求鸡兔各多少只？

按照常规求解方法我们会利用一元二次方程，假设鸡的数量为X_1，而兔的数量为X_2，则这个问题可以用以下两个一元二次方程表示：

$$X_1 + X_2 = 35$$

$$2X_1 + 4X_2 = 94$$

使用一元二次方程然可以迅速求解出结果，如果使用Excel的规划求解工具，操作步骤如下。

（1）建立Excel模型

将题目条件在Excel中建立模型，如图2-3所示。

	A	B	C
1	鸡数量		
2	兔数量		
3	总头数	0	35
4	总脚数	0	94

图2-3　求解模型

其中B1与B2单元格是空单元格，为未知数值，即需要求解的X_1和X_2。

B3单元格中为总头数的计算公式，即：

$$X_1+X_2=35$$

B4单元格中为总脚数的计算公式，即：

$$2X_1+4X_2=94$$

（2）规划求解设置

在Excel菜单栏中单击【数据】|【规划求解参数】选项。显示【规划求解参数】对话框，如图2-4所示。



图2-4　【规划求解参数】选项

在【遵守约束】下拉列表框中添加约束条件，单击【添加】按钮。在【通过更改可变单元格】下拉列表框中选择B1:B2单元格区域，在【遵守约束】下拉列表框中输入条件1"B3:B4＝C3:C4”，即满足方程式右边的条件；输入条件2"B1:B2＝整数”。单击【确定】按钮，返回【规划求解参数】对话框。

（3）查看分析结果

选择【单纯线性规划】选项，单击【求解】按钮求解，结果如图2-5所示。

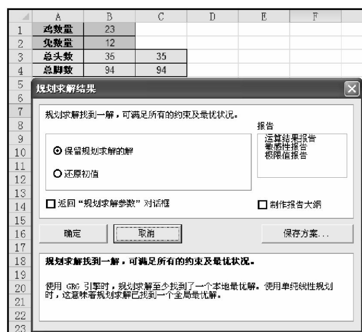


图2-5　求解结果

求解出鸡的数量为23只，兔的数量为12只。选择【保存方案】按钮，单击【确定】按钮。

Miss Ju：“Excel还可以这样用啊，真的是太神奇了，可是我现在还不是太明白其求解原理。”

Mr Shu：“其实原理也不复杂，无非是利用运筹学中的线性规划原理。不用着急，先知道原理。我们再列举几个例子，多用几个软件来求解就会明白。”

案例：**Excel**规划求解——生产组合

这是一个简单的生产组合问题，某家工厂生产某系列的3种型号的产品。这3种产品的使用原材料相同，但材料消耗量和产品市场销售价格有所不同，目前工厂的原材料库存、销售价格和最低生产要求如图2-6所示。

	A	B	C	D
1	现有原料	5000		
2				
3		原料消耗	销售价格	最低要求产量
4	型号A	3	20	300
5	型号B	2	13	400
6	型号C	4	25	280

图2-6　产品消耗信息

说明如下。

·生产A产品需要消耗3个单位的原材料，售价20，最低产量要求300。

·生产B产品需要消耗2个单位的原材料，售价13，最低产量要求400。

·生产C产品需要消耗4个单位的原材料，售价25，最低产量要求280。

这是典型的线性规划问题，每种产品消耗的原材料和销售价格不同。要想获取最大的产值，必须考虑约束条件限制。即每种产品有最低产量限制，总原材料数也是限制条件。如果按照常规思路求解，需要进行多次迭代试运算。应用Excel的规划求解工具可以迅速求解，操作步骤如下。

（1）建立Excel模型

将已知条件在Excel中建立模型，打开生产组合-规划求解.xls工作表，模型设置如下。

E3:G3区域用于存放求解变量，标题分别设置为“目标产量”、“原料消耗量”和“目标产值”。

设置原料消耗量计算公式，在F4单元格内输入公式“＝$E*B4”。复制F4单元格的公式向下填充至F6单元格，在F7单元格内求和。

同理设置目标产值公式，计算公式为产量×销售价格。在G7单元格为产值求和，结果如图2-7所示。

	A	B	C	D	E	F	G
1	现有原料	5000					
2							
3		原料消耗	销售价格	最低要求产量	目标产量	原料消耗	目标产值
4	型号A	3	20	300	0	0	0
5	型号B	2	13	400	0	0	0
6	型号C	4	25	280	0	0	0
7	合计					0	0

图2-7　生产组合模型

（2）设置规划求解

在Excel菜单栏中单击【数据】|【规划求解】选项，显示【设置

规划求解参数】对话框。设置参数，如图2-8所示。



图2-8　设置规划求解参数

在【设置目标】下拉列表框内选择G7单元格，选择【最大值】单选按钮，即目标是产值最大化。

在【通过更改可变单元格】下拉列表框中选择E4:E6单元格区域用于保存需要求解的答案， 即每种产品的产量。

在【遵守约束】下拉列表框中内添加约束条件，单击【添加】按钮，设置限制条件如下。

- ·条件1：3种产品分别有最低产量限制，输入E4＞＝G4、E5＞＝G5和E6＞＝G6。
- ·条件2：原料消耗要求，总的生产消耗不能超过目前的原料存量，因此F7＜＝B1。
- ·条件3：产量整数要求，根据实际情况而定。在本例中需要生产的产品需要是整数，输入E4:E6＝整数。

在添加整数要求时需要在条件格式中选择【int】选项，如图2-9

所示。



图2-9　添加整数要求

设置完成后单击【确定】按钮，返回【规划求解参数】对话框。

（3）查看规划求解结果

在【规划求解参数】对话框中选择【单纯线性规划】选项。单击【求解】按钮，结果如图2-10所示。

	A	B	C	D	E	F	G
1	现有原料	5000					
2							
3		原料消耗	销售价格	最低要求产量	目标产量	原料消耗总量	目标产值
4	型号A	3	20	300	1026	3078	20520
5	型号B	2	13	400	401	802	5213
6	型号C	4	25	280	280	1120	7000
7	合计					5000	32733

规划求解结果
规划求解找到一解，可满足所有的约束及最优状况。
报告
运算结果报告
保留规划求解的解
还原初值
返回“规划求解参数”对话框
制作报告大纲
确定
取消
保存方案...
规划求解找到一解，可满足所有的约束及最优状况。
使用 GRG 引擎时，规划求解至少找到了一个本地最优解。使用单纯线性规划时，这意味着规划求解已找到一个全局最优解。

图2-10　生产组合求解结果

计算结果为A产品生产1　027吨，消耗原材料3　078；B产品生产400吨，消耗原材料802；C产品生产280吨，消耗原材料1 120。

目标函数产值最大为32　733，原料5　000吨已全部消耗完。从计算结果来看更趋向于生产A产品，尽管其销售价格不是最高。

求解之后可以单击【保存方案】按钮，单击【确定】按钮。或关闭【规划求解结果】对话框，并将结果保留在Excel表中。如果选中

【恢复初值】单选按钮或单击【取消】按钮，Excel表恢复到使用【规划求解】之前的状态。

案例：Excel规划求解——油品勾兑配比

石油精炼是国民经济的基础性行业，在石油精炼过程中通常会使用勾兑。某工厂从事原料的加工，外购5种原油来加工生产成品油种。

生产的成品油主要有甲、乙、丙和丁4种主要成分及少量其他杂质，5种主要原油原料来自5个供应商。他们所提供的原油中甲、乙、丙和丁4种成分的含量各不相同，原油的价格也有所区别。整理各供应商的原料的成分及价格数据并保存在“配料问题”工作表中，如图2-11所示。

	A	B	C	D	E	F
1		成分含量(百分比)				
2		甲	乙	丙	丁	原油价格(每吨)
3	原油1	25%	8%	30%	29%	5350
4	原油2	30%	12%	38%	14%	5370
5	原油3	45%	9%	21%	18%	5390
6	原油4	38%	7%	32%	16%	5393
7	原油5	33%	10%	28%	20%	5362

图2-11　各供应商的原料的成分及价格

产品质量标准的要求为单元成品油中必须包含甲成分不少于36%，乙成分不多于12%，丙成分在28%～35%之间，丁成分不少于18%，其他杂质不多于7%。

工厂生产计划部门要求计算如何选择5种原有原料进行配比可以使得产品满足质量要求的情况下成本最低？各种原油之间的组分差异较大，但原油价格差异并不是很明显。

这也是线性规划的典型应用，目标为成本最低，约束条件为成品油的组分含量。此时使用方程求解已经不能满足需求，使用Excel【规划求解】求解的操作步骤如下。

（1）建立Excel模型

需要计算的变量是每种原油的使用量，以1吨成品油为计算单位。根据使用量计算混合之　　后的各个组分含量满足成品油的要求即可，在Excel中建立的模型如图2-12所示。

	A	B	C	D	E	F	G
1		成分含量(百分比)					
2		甲	乙	丙	丁	原油价格(每吨)	
3	原油1	25%	8%	30%	29%	5350	
4	原油2	30%	12%	38%	14%	5370	
5	原油3	45%	9%	21%	18%	5390	
6	原油4	38%	7%	32%	16%	5393	
7	原油5	33%	10%	28%	20%	5362	
8							
9							
10		甲	乙	丙	丁	实际使用数量(吨)	原料成本
11	原油1	0	0	0	0	0	0
12	原油2	0	0	0	0	0	0
13	原油3	0	0	0	0	0	0
14	原油4	0	0	0	0	0	0
15	原油5	0	0	0	0	0	0
16	合计	0%	0%	0%	0%	0	0
17	杂质含量	100%					

图2-12　建立的模型

在Excel中设置如下。

·F11:F15单元格区域为实际的各原料用量，作为规划求解的可变单元格。

·B11:E16单元格区域使用公式计算每种成分的含量，在B11单元中输入“＝$F11*B3”，其他以此类推；B16:E16使用求和公式完成；B17单元格表示杂质的含量，输入公式“＝1-XUM(B16:E16)”。

（2）设置规划求解

在Excel菜单中单击【数据】|【规划求解】选项，显示【规划求解参数】对话框。在【通过更改可变单元格】下拉列表框中选择G16单元格保存1吨成品油的总成本。选中【最小值】单选按钮，即成本最低。

在【遵守约束】下拉列表框内添加约束条件，单击【添加】按钮，设置如下。

·条件1：每种原料油需为非负，输入F11:F15＞＝0。

·条件2：甲成分不少于36%，输入B16＞＝0.36。

·条件3：乙成分不多于12%，输入C16＜＝0.12。

·条件4：丙成分不少于28%，输入D16＞＝0.28。

·条件5：丙成分不多于35%，输入D16＜＝0.35。

·条件6：丁成分不少于18%，输入E16＞＝0.18。

·条件7：杂质不多于7%，输入B17＜＝0.07。

·条件8：混合总量＝1，输入F16＝1。

设置完成后如图2-13所示。



图2-13　设置规划求解参数

（3）求解模型

单击【求解】按钮，计算结果如图2-14所示。

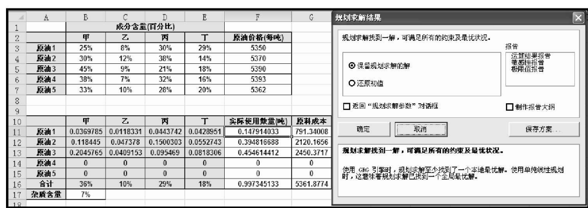


图2-14　规划求解计算结果

在满足所有约束条件下1吨成品油需要使用5种原料的数量分别如下。

·原油1：0.148吨。

·原油2：0.395吨。

·原油3：0.455吨，其他两种油品不需要。

1吨成品油的最低成本为5361.8元。

（4）解读计算结果

按住Ctrl键选择【规划求解结果】对话框【报告】列表框中的【运算结果报告】、【敏感性报告】和【极限值报告】选项，如图2-15所示。

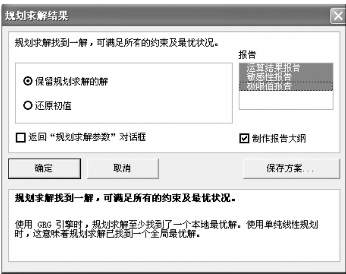


图2-15　选择3个选项

单击【确定】按钮，显示的求解报告如图2-16所示。

目标单元格（最小值）

单元格	名称	初值	终值
$G$16	合计 原料成本	0	5361.87737

可变单元格

单元格	名称	初值	终值	整数
$F$11	原油1 实际使用数量(吨)	0	0.147914033	约束
$F$12	原油2 实际使用数量(吨)	0	0.394816688	约束
$F$13	原油3 实际使用数量(吨)	0	0.454614412	约束
$F$14	原油4 实际使用数量(吨)	0	0	约束
$F$15	原油5 实际使用数量(吨)	0	0	约束

约束

单元格	名称	单元格值	公式	状态	型数值
$B$16	合计 甲	36%	$B$16>=0.36	到达限制值	0%
$B$17	杂质含量 甲	7%	$B$17<=0.07	到达限制值	0
$C$16	合计 乙	10%	$C$16<=0.12	未到限制值	0.019873578
$F$16	合计 实际使用数量(吨)	0.997345133	$F$16<=1	未到限制值	0.002654867
$D$16	合计 丙	29%	$D$16>=0.28	未到限制值	1%
$D$16	合计 丙	29%	$D$16<=0.35	未到限制值	0.060126422
$E$16	合计 丁	18%	$E$16>=0.18	到达限制值	0%
$F$11	原油1 实际使用数量(吨)	0.147914033	$F$11>=0	未到限制值	0.147914033
$F$12	原油2 实际使用数量(吨)	0.394816688	$F$12>=0	未到限制值	0.394816688
$F$13	原油3 实际使用数量(吨)	0.454614412	$F$13>=0	未到限制值	0.454614412
$F$14	原油4 实际使用数量(吨)	0	$F$14>=0	到达限制值	0
$F$15	原油5 实际使用数量(吨)	0	$F$15>=0	到达限制值	0

图2-16　求解报告

Miss Ju：“你一口气列举了3个Excel规划求解的例子，以前不知道Excel还有这个功能。规划求解功能真的太强大了，不过好像还蛮好理解。”

Mr　Shu：“是的，我们一般情况下都只用Excel的最基本的功能，类似规划求解这种工具是数据分析的利器。”

Miss　Ju：“你前面提到规划求解是较强大的数据分析功能，那么除此之外还有哪些数据分析功能呢？”

Mr　Shu：“是的，其实除了规划求解该功能以外，Excel在数据分析方面还有一些其他功能，我们通过下面几个例子来说明。”

案例：Excel理财——模拟运算表

模拟运算表是假设分析的方法之一，它可以显示公式中一个或两个参数的变动对计算结果的影响。从而求得某一运算中可能发生的数值变化，并将这些变化列在一个表中以便比较。由于观察的变量多少不同，因此分为单变量模拟运算表和双变量模拟运算表两种形式，前者多用于财务金融分析。

现有一投资案例，假设初始投资5万元，年收益率假定为7%，投资时间为20年，要求测算到期的资金总额及相关收益。这相当于在银行存定期，根据各计算元素之间的关系建立表格。

（1）建立初始模型

到期总额（本金+利息)=$5\times(1+i)^{n}$（5为初始本金，i 为利息，n 为年数），收益=到期总额-本金，收益率=收益/本金。打开“单变量模拟运算表”工作表，在Excel中模拟，结果如图2-17所示。

	A	B
1	初始投资	50000
2	年收益率	7%
3	投资时间(年)	20
4		
5	到期总额	193484.22
6	总收益	143484.22
7	总收益率	287%

图2-17　模拟结果

这是一个非常简单的模型，模型设置为所有已知的初始条件分别位于B1:B3单元格中，B5:B7单元格区域保存得到的结果。在B5单元格中输入公式“=B1*(1+B2)^B3”，在B6单元格中输入公式“=B5-B1”，

在B7单元格中输入公式“＝B6/B1”即可得到我们想要的结果。

（2）修订模型

在现实情况中已知的一些条件往往无法确定，如年收益率可能是一个变化的利率，不是7%；投资时间也不会是20年，也许我们想知道的是10年～15年的收益。上面的公式只能让我们知道在年收益率7%且投资时间为20年这种情况下的收益为14.3万，其他变化情况仅凭上面一组公式无法知道答案，需要利用公式计算。现在我们要测算年收益率从3%～7%且每增加0.5%，20年后到期总金额的变化情况。

设置如下。

·在D1:L1单元格区域从3%每隔0.5%填充至7%。

·在D2单元格中输入公式“$B$1*(1＋B2)Λ$B$3”，注意公式需要对B1和B3单元格地址绝对引用。

·将D2单元格中的公式向右复制至L2。

·将收益率的公式在D3～L3中复制，结果如图2-18所示。

	A	B	C	D	E	F	G	H	I	J	K	L
1	初始投资	50000		3.0%	3.5%	4.0%	4.5%	5.0%	5.5%	6.0%	6.5%	7.0%
2	年收益率	7%		90305.562	99489.443	109556.16	120585.7	132664.89	145887.87	160356.77	176182.25	193484.22
3	投资时间(年)	20		80.6%	99.0%	119.1%	141.2%	165.3%	191.8%	220.7%	252.4%	287.0%
4												
5	到期总额	193484.22										
6	总收益	143484.22										
7	总收益率	287%										

图2-18　修订后的收益模型

以上利用公式函数的方式计算收益，这是最原始和初级的方法。这时模拟计算表很方便，下面我们就用其来执行上述运算。

（3）单变量模拟计算（年收益率）

在整个投资计算模型中对收益有影响的主要因素有3项，分别为初始投资、年收益率和投资时间，其中任意一项因素的变化都可以影响最终结果。如果计算模型中只有一个变量因素在发生变化，我们可以称之为“单变量模拟运算表”。

操作步骤如下。

·在D5～L5单元格中从3%每隔0.5%填充至7%。

·在C6单元格内输入公式“＝B5”表示引用该公式。

·选中C5:L6，单击【数据】|【模拟分析】|【模拟运算表】选项。在【模拟运算表】对话框中【输入应用行的单元格】下拉列表框中输入“$B$2”，如图2-19所示。

	A	B	C	D	E	F	G	H	I	J	K	L
1	初始投资	50000		3.0%	3.5%	4.0%	4.5%	5.(				7.0%
2	年收益率	7%		90305.562	99489.443	109556.16	120585.7	1326				93484.22
3	投资时间(年)	20		80.6%	99.0%	119.1%	141.2%	165				287.0%
4												
5	到期总额	193484.22		3.0%	3.5%	4.0%	4.5%	5.(				7.0%
6	总收益	143484.22	193484.22									
7	总收益率	287%										
8												

模拟运算表
输入引用行的单元格(R)：$B$2
输入引用列的单元格(C)：
确定 取消

图2-19　设置模拟运算表

·单击【确定】按钮，计算结果如图2-20所示。

	A	B	C	D	E	F	G	H	I	J	K	L
1	初始投资	50000		3.0%	3.5%	4.0%	4.5%	5.0%	5.5%	6.0%	6.5%	7.0%
2	年收益率	7%		90305.562	99489.443	109556.16	120585.7	132664.89	145887.87	160356.77	176182.25	193484.22
3	投资时间(年)	20		80.6%	99.0%	119.1%	141.2%	165.3%	191.8%	220.7%	252.4%	287.0%
4												
5	到期总额	193484.22		3.0%	3.5%	4.0%	4.5%	5.0%	5.5%	6.0%	6.5%	7.0%
6	总收益	143484.22	193484.22	90305.562	99489.443	109556.16	120585.7	132664.89	145887.87	160356.77	176182.25	193484.22
7	总收益率	287%										
8												

图2-20　计算结果

结果与用公式计算一样。如果需要查看变量“年收益率”变化对总收益的影响，同样可以采用上述方法。要使模拟运算表能够计算，必须具备以下几个条件。

·运算公式：公式的作用是告知Excel如何从参数得到用户希望了解的结果，可以直接输入。例如，上述例子中可以在C6中直接输

入公式“$B$1*(1＋B2)Λ$B$3”，也可以输入公式所在的单元格地址。

·工作区域：包括变化参数、运算公式及运算结果的存放位置，其中变化参数的存放位置默认为模拟运算表工作区域的首行或首列，这里的首行是模拟运算表的参数引用行；首列为模拟运算表的参数引用列。

·变量：变量指整个模拟运算表中所需要调整变化的参数，也是整个运算的重要依据，本例中收益率是变量。

如果假设“年收益率”参数固定为7%，而参数“投资时间”在10年～20年之间波动，计算每年到期的总金额变化情况。

（4）单变量（投资时间）模拟计算

·选中D9:D19单元格区域，依次输入10年～20年每隔1年的各个时间点，在E8单元格内输入公式“＝$B$1*(1＋B2)Λ$B$3”。

·选定单元格区域D8:E19，选择【数据】|【模拟分析】|【模拟运算表】选项。在【模拟运算表】对话框的【输入引用列的单元格】下拉列表框中选中“$B$3”单元格，如图2-21所示。



图2-21　设置模拟运算表

·单击【确定】按钮，计算结果如图2-22所示。

	A	B	C	D	E	F	G	H	I	J	K	L
1	初始投资	50000		3.0%	3.5%	4.0%	4.5%	5.0%	5.5%	6.0%	6.5%	7.0%
2	年收益率	7%		90305.562	99489.443	109556.16	120585.7	132664.89	145887.87	160356.77	176182.25	193484.22
3	投资时间(年)	20		80.6%	99.0%	119.1%	141.2%	165.3%	191.8%	220.7%	252.4%	287.0%
4												
5	到期总额	193484.22		3.0%	3.5%	4.0%	4.5%	5.0%	5.5%	6.0%	6.5%	7.0%
6	总收益	143484.22	193484.22	90305.562	99489.443	109556.16	120585.7	132664.89	145887.87	160356.77	176182.25	193484.22
7	总收益率	287%										
8					193484.22							
9				10	98357.568							
10				11	105242.6							
11				12	112609.58							
12				13	120492.25							
13				14	128926.71							
14				15	137951.58							
15				16	147608.19							
16				17	157940.76							
17				18	168996.61							
18				19	180826.38							
19				20	193484.22							

图2-22　计算结果

参数“投资时间”不仅影响到期总额，还会影响总收益和收益率等，利用模拟运算表可以 快速计算。如果还想知道参数“投资时间”对总收益和总收益率的影响，只需要在上述步骤中将运算公式和工作区域进行扩展即可。将F8单元格引用B6单元格公式，将G8单元格引用F8单元格公式，工作区域选择D8:G19，单击【确定】按钮，计算结算如图2-23所示。

	A	B	C	D	E	F	G	H	I	J	K	L
1	初始投资	50000		3.0%	3.5%	4.0%	4.5%	5.0%	5.5%	6.0%	6.5%	7.0%
2	年收益率	7%		90305.562	99489.443	109556.16	120585.7	132664.89	145887.87	160356.77	176182.25	193484.22
3	投资时间(年)	20		80.6%	99.0%	119.1%	141.2%	165.3%	191.8%	220.7%	252.4%	287.0%
4												
5	到期总额	193484.22		3.0%	3.5%	4.0%	4.5%	5.0%	5.5%	6.0%	6.5%	7.0%
6	总收益	143484.22	193484.22	90305.562	99489.443	109556.16	120585.7	132664.89	145887.87	160356.77	176182.25	193484.22
7	总收益率	287%										
8												
9				10	98357.568	48357.568	96.7%					
10				11	105242.6	55242.598	110.5%					
11				12	112609.58	62609.579	125.2%					
12				13	120492.25	70492.25	141.0%					
13				14	128926.71	78926.708	157.9%					
14				15	137951.58	87951.577	175.9%					
15				16	147608.19	97608.187	195.2%					
16				17	157940.76	107940.76	215.9%					
17				18	168996.61	118996.61	238.0%					
18				19	180826.38	130826.38	261.7%					
19				20	193484.22	143484.22	287.0%					

图2-23　计算结果

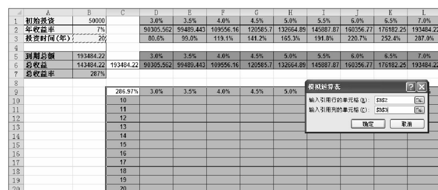
在单变量求解过程中均是单个变量在变化，如果“年收益率”和“投资时间”两个变量均在变化且需要计算“总收益率”的变化，通过公式固然可以实现，但比较烦琐。通过模拟运算表则相对比较容易，此时需要执行双变量模拟计算。

（5）双变量模拟运算

在Excel中建立模型，操作步骤如下。

·将C9单元格引用B7单元格的公式或直接输入B7单元格中的公式“＝(B1*(1＋B2)^B3-B1)/B1”。

·选择单元格区域D9:G19，选择【数据】|【模拟分析】|【模拟运算表】选项，显示【模拟运算表】对话框。在【输入行引用单元格】下拉列表框中输入“$B$2”，在“输入列引用单元格”下拉列表框中输入“$B$3”，如图2-24所示。



	A	B	C	D	E	F	G	H	I	J	K	L
1	初始投资	50000		3.0%	3.5%	4.0%	4.5%	5.0%	5.5%	6.0%	6.5%	7.0%
2	年收益率	7%		90305.562	99489.443	109556.16	120585.7	132664.89	145887.87	160356.77	176182.25	193484.22
3	投资时间(年)	20		80.6%	99.0%	119.1%	141.2%	165.3%	191.8%	220.7%	252.4%	287.0%
4												
5	到期总额	193484.22		3.0%	3.5%	4.0%	4.5%	5.0%	5.5%	6.0%	6.5%	7.0%
6	总收益	143484.22	193484.22	90305.562	99489.443	109556.16	120585.7	132664.89	145887.87	160356.77	176182.25	193484.22
7	总收益率	287%										
8												
9			286.97%	3.0%	3.5%	4.0%	4.5%	5.0%				
10			10									
11			11									
12			12									
13			13									
14			14									
15			15									
16			16									
17			17									
18			18									
19			19									
20			20									

图2-24　设置双变量模拟运算

·单击【确定】按钮，计算结果如图2-25所示。

	A	B	C	D	E	F	G	H	I	J	K	L
1	初始投资	50000		3.0%	3.5%	4.0%	4.5%	5.0%	5.5%	6.0%	6.5%	7.0%
2	年收益率	7%		90305.562	99489.443	109556.16	120585.7	132664.89	145887.87	160356.77	176182.25	193484.22
3	投资时间(年)	20		80.6%	99.0%	119.1%	141.2%	165.3%	191.8%	220.7%	252.4%	287.0%
4												
5	到期总额	193484.22		3.0%	3.5%	4.0%	4.5%	5.0%	5.5%	6.0%	6.5%	7.0%
6	总收益	143484.22	193484.22	90305.562	99489.443	109556.16	120585.7	132664.89	145887.87	160356.77	176182.25	193484.22
7	总收益率	287%										
8												
9			286.97%	3.0%	3.5%	4.0%	4.5%	5.0%	5.5%	6.0%	6.5%	7.0%
10			10	34.4%	41.1%	48.0%	55.3%	62.9%	70.8%	79.1%	87.7%	96.7%
11			11	38.4%	46.0%	53.9%	62.3%	71.0%	80.2%	89.8%	99.9%	110.5%
12			12	42.6%	51.1%	60.1%	69.6%	79.6%	90.1%	101.2%	112.9%	125.2%
13			13	46.9%	56.4%	66.5%	77.2%	88.6%	100.6%	113.3%	126.7%	141.0%
14			14	51.3%	61.9%	73.2%	85.2%	98.0%	111.6%	126.1%	141.5%	157.9%
15			15	55.8%	67.5%	80.1%	93.5%	107.9%	123.2%	139.7%	157.2%	175.9%
16			16	60.5%	73.4%	87.3%	102.2%	118.3%	135.5%	154.0%	173.9%	195.2%
17			17	65.3%	79.5%	94.8%	111.3%	129.2%	148.5%	169.3%	191.7%	215.9%
18			18	70.2%	85.7%	102.6%	120.8%	140.7%	162.1%	185.4%	210.7%	238.0%
19			19	75.4%	92.3%	110.7%	130.8%	152.7%	176.6%	202.6%	230.9%	261.7%
20			20	80.6%	99.0%	119.1%	141.2%	165.3%	191.8%	220.7%	252.4%	287.0%

图2-25　计算结果

模拟运算相对于其他数据分析工具较为简单，但只要能灵活运用，有时也能达到事半功倍的效果。使用模拟运算表同样可以解决“鸡兔同笼”问题，具体操作方法读者可以思考。

案例：Excel变身大管家——方案管理器

模拟运算表只能分析公式中一个或两个参数的变化对计算结果的影响，但在实际工作中往往会遇到更多的变动因素，Excel数据分析工具中还有“方案管理器”可以来解决此类问题。例如，在前面的双变量模拟运算表中我们知道“年收益率”和“投资时间”两个变量对“总收益

率”的影响，而并没有考虑“初始投资”对“总收益率”的影响，此时使用“方案管理器”这一工具就可以轻松实现。

为了增强方案的可读性，需要用到Excel中的“名称”的概念，它可以理解为单元格内容的名字。名称是Excel高级操作中的常用功能，这里我们只介绍其数据分析功能。打开“方案管理器”工作表，操作步骤如下。

（1）定义名称

单击B1单元格，在名称框中输入“初始投资”。单击B2单元格，在名称框中输入“年收益率”。其他依次类推，如图2-26所示。

初始投资 fx

	A	B
1	初始投资	50000
2	年收益率	7%
3	投资时间(年)	20
4	名称框	
5	到期总额	193484.22
6	总收益	143484.22
7	总收益率	287%

图2-26　设置名称

（2）打开方案管理器

在Excel菜单中选择【数据】|【模拟分析】|【方案管理器】选项，显示如图2-27所示的【方案管理器】对话框。

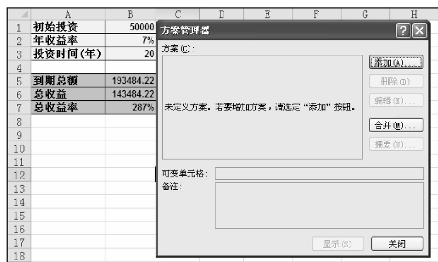


图2-27　【方案管理器】对话框

（3）添加方案

在管理器中添加长期、中期和短期3种方案。

·单击【添加】按钮，显示【编辑方案】对话框。在【可变单元格】下拉列表框内选择B1:B3单元格区域，表示可以变化变量的区域。在【方案名】文本框中输入“长期投资方案”，如图2-28所示。

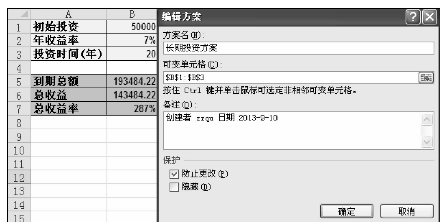


图2-28　编辑方案

·单击【确定】按钮，显示【方案变量值】对话框，如图2-29所示。



图2-29 【方案变量值】对话框

·在【初始投资】文本框中输入“50 000”，在【年收益率】文本框中输入“0.07”，在【投资时间_年】文本框中输入“20”，这几个名称就是刚才我们在前面定义的名称。

·单击【添加】按钮，重复以上的操作分别定义短期和中期的投资方案，参数分别设置如表2-1所示。

表**2-1** **3**种投资方案

方　　案	初始投资（元）	年收益率（%）	投资时间（年）
长期投资方案	50 000	0.07	20
中期投资方案	300 000	0.05	15
短期投资方案	200 000	0.03	3

·添加完成后的【方案管理器】对话框如图2-30所示。

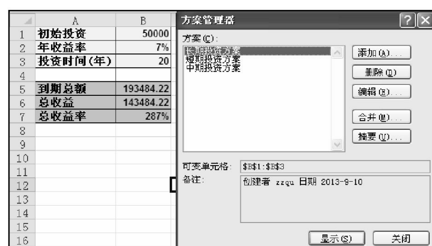


图2-30 添加完成后的【方案管理器】对话框

（4）各方案的计算结果

选中想要显示的方案的结果，单击【显示】按钮，计算结果如图2-31所示。

图2-31　计算结果

（5）获取方案报告

在【方案管理器】对话框中选择任意一个方案，单击【摘要】按钮，如图2-32所示。

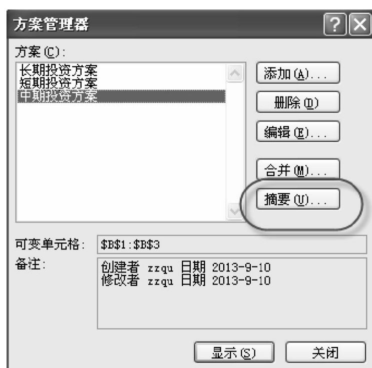


图2-32　【摘要】按钮

显示【方案摘要】对话框，其中包括【方案摘要】与【方案数据透视表】两个选项，前者会以大纲形式展示方案报告；后者会以数据透视表形式展示报告。选择后在【结果单元格】下拉列表框中选择B5:B7单元格区域，如图2-33所示。

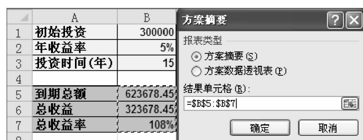


图2-33　选择结果单元格

单击【确定】按钮，Excel会自动创建一个新的工作表并且命名为“方案摘要”；同时激活这个工作表显示所生成的摘要报告，如图2-34所示。

方案摘要				
	当前值：	长期投资方案	短期投资方案	中期投资方案
可变单元格：				
初始投资	300000	50000	200000	300000
年收益率	5%	7%	3%	5%
投资时间_年	15	20	3	15
结果单元格：				
到期总额	623678.4538	193484.2231	218545.4	623678.4538
总收益	323678.4538	143484.2231	18545.4	323678.4538
总收益率	108%	287%	9%	108%

注释：“当前值”这一列表示的是在
建立方案汇总时，可变单元格的值。
每组方案的可变单元格均以灰色底纹突出显示。

图2-34　摘要报告

如果在上一步操作中选择【方案数据透视表】单选按钮，Excel同样会自动生成一个名为“方案数据透视表”的工作表，如图2-35所示。

	A	B	C	D
1	$B$1:$B$3 由	(全部)		
2				
3		结果单元格		
4	$B$1:$B$3	到期总额	总收益	总收益率
5	长期投资方案	193484.2231	143484.2231	287%
6	短期投资方案	218545.4	18545.4	9%
7	中期投资方案	623678.4538	323678.4538	108%
8				
9				
10				
11				
12				
13				
14				
15				

数据透视表字段列表
选择要添加到报表的字段：
$B$1:$B$3
$B$1:$B$3 由
res 到期总额
res 总收益
res 总收益率
在以下区域间拖动字段：
报表筛选　$B$1:$B$3 由
列标签　Σ 数值
行标签　$B$1:$B$3
Σ 数值　到期总额
推迟布局更新　更新

图2-35　数据透视表

从本质上说，方案管理器的运行实际就是将方案中设置保存的变量值赋给对应的变量单元格，通过工作表中已有的这些变量的公式来得到相应的结果。

案例：重新数鸡——综合应用方案管理器与模拟运算表

前面我们利用规划求解数过鸡，用模拟运算及方案管理器理过财。

在利用规划求解计算“鸡兔同笼”问题的过程中鸡兔的总数只有一个。如果鸡兔的总数是一组数据，则使用规划求解可能要建立多个模型，我们尝试结合方案管理器及模拟运算表进行求解。

需要求解的数据如表2-2所示。

表**2-2**　需要求解的数据组

组　　数	总　头　数	总　脚　数
第一题	35	94
第二题	25	88
第三题	30	66
第四题	49	150
第五题	38	138

此时单独使用模拟运算表或方案管理器已经不能很方便地解决问题，将两项工具结合使用则可以很好解决。打开“鸡兔同笼”工作表，操作步骤如下。

（1）定义名称

定义A2单元格的名称为“总头数”，B2单元格定义为“总脚数”，A4单元格定义为“兔子个数”，B4单元格定义为“鸡个数”。如图2-36所示。

总头数

	A	B
1	总头数	总脚数
2	38	138
3	兔子个数	鸡个数
4	31	7

图2-36　定义名称

（2）添加方案管理器

·选择【数据】|【模拟分析】|【方案管理器】选项，显示【编辑方案】对话框，如图2-37所示。

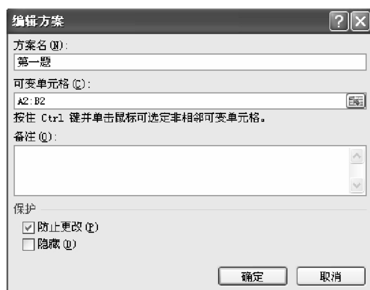


图2-37　【编辑方案】对话框

·在【方案名】文本框中输入“第一题”，在【可变单元格】下拉列表框中输入“A2:B2”单元格区域。

·单击【确定】按钮，显示【方案变量值】对话框，如图2-38所示。



图2-38　【方案变量值】对话框

在【总头数】文本框中输入“35”，在【总脚数】文本框中输入“94”。

·单击【确定】按钮，添加完成第1个方案。

·重复以上步骤，直至完成添加其他方案，如图2-39所示。



图2-39　添加方案

（3）获取方案报告

单击【摘要】按钮，显示【方案摘要】对话框中。在【结果单元格】中选择A4:B4单元格区域，单击【确定】按钮，求解结果如图2-40所示。

方案摘要	当前值:	第一题	第二题	第三题	第四题	第五题
可变单元格:						
总头数	38	35	25	30	49	38
总脚数	138	94	88	66	150	138
结果单元格:						
兔子个数	31	12	19	3	26	31
鸡个数	7	23	6	27	23	7

注释：“当前值”这一列表示的是在建立方案汇总时，可变单元格的值。
每组方案的可变单元格均以灰色底纹突出显示。

图2-40　求解结果

该例是模拟运算表与方案管理器结合的典型应用，前者根据变化的参数值计算正确的结果；后者提供参数的变化值，并且显示和汇总结果。

案例：Excel逆运算——单变量求解

Excel中还有一个常用的数据分析功能，即单变量求解。它是对某一功能问题按公式计算所得结果做出的假设，以推测公式中形成结果的一系列变量可能发生的变化。它是由“果”求“因”的逆运算，从方程的角度来讲，即已知Y 求解X 。

仍以上面的简单投资为例进行说明，在已知“初始投资”、“年收益率”和“投资时间”3个参数的条件下可以计算出“到期总额”、“总收益”和“总收益率”等。假设某个用户想了解要在20年内将“总收益率”提升至200%，至少要保证每年多少的“年收益率”才能达到这个目标？很显然，这就是典型的逆向分析，即通过结果求取条件。单变量求解适用这种需求，打开“单变量求解”工作表，操作步骤如下。

（1）建立模型

模型与之前保持一致，如图2-41所示。

	A	B
1	初始投资	50000
2	年收益率	7.000%
3	投资时间(年)	20
4		
5	到期总额	193484.22
6	总收益	143484.22
7	总收益率	287%

图2-41　数据模型

（2）单变量求解

选择【数据】|【模拟分析】|【单变量求解】选项，显示【单变量求解】对话框。在【目标 单元格】下拉列表框中选择B7单元格，即总收益率。在【目标值】文本框中输入300%，在【可变单元格】下拉列表框中选择需要求解的内容，即“年收益率”的B2单元格，如图2-42所示。

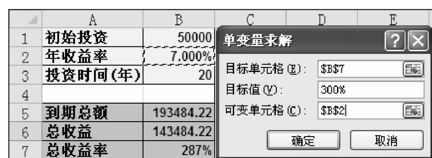


图2-42　设置单变量求解参数

单击【确定】按钮，求解结果如图2-43所示。

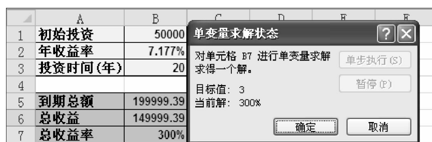


图2-43　单变量求解结果

从以上计算可以看出在逆向计算时用户只需要建立正向模型，然后利用单变量求解即可。

Miss Ju：“哇，真没想到Excel还有这么多的功能！”

Mr　Shu：“其实Excel的数据分析功能还远不止这些，它还有强大的数据透视表和VBA。”

Miss　Ju：“这么好，我对Excel数据分析功能中最感兴趣的还是规划求解，它太神奇了。”

Mr Shu：“是的，线性规划基本上基于运筹学的线性规划的原理，可提供与Excel规划求解类似功能的还有很多软件，如我们前面一直在说的Crystal　Ball、Minitab和LINGO。Minitab做优化分析不是特别方便，但是CB和LINGO两款软件可以说是优化分析中的佼佼者。如果能掌握这两款软件的优化分析功能，其他很多软件的功能也就大同小异了，如@Risk等。”

Miss Ju：“但对于优化分析的一些概念我还不是特别理解，如线性规划等，你能简单讲讲它的原理吗？”

Mr Shu：“好的，先了解一些概念也有助于我们理解软件的操作，在了解它的原理之后可以再看看Crystal Ball和LINGO的威力。”

2.2 做一个精明的农场主

线性规划是运筹学中研究较早、发展较快、应用广泛和方法较成熟的一个重要分支，它是辅助人们科学管理的一种数学方法。在经济管理、交通运输和工农业生产等经济活动中提高经济效益是人们不可缺少的要求，而提高经济效益一般通过两种途径：一是技术方面的改进，如改善生产工艺，使用新设备和新型原材料；二是生产组织与计划的改进，即合理安排人力物力资源。线性规划所研究的是在一定条件下合理安排人力物力等资源，使经济效益达到最好。一般地求线性目标函数在线性约束条件下的最大值或最小值的问题统称为“线性规划问题”。满足线性约束条件的解叫做“可行解”，由所有可行解组成的集合叫做“可行域”。

线性规划（Linear Programming，LP）方法为解决静态确定性决策中的问题提供了有效的途径，本节介绍线性规划模型的基本概念，并且讨论仅有两个决策变量的LP问题的几何解法，其目的是深入理解和掌握LP问题的基本原理。目标函数（预测变量）、约束条件和决策变量是线性规划的3个要素。

（1）目标函数

目标函数是一个寻求最大值、最小值或特定值的线性数学表达式，表明一个使某个变量最大或最小或特定值的目的。

（2）约束条件（简写为“S.T “）

使用有效资源的限制条件。

（3）决策变量

决策者可以控制的因素，其值决定目标函数的值。

构成LP模型的必要条件如下。

·目标明确：所有的问题都要求解某个最大值或最小值，称为“目标函数”，该函数应是一组决策变量的线性函数。

·资源有限：LP问题是有约束的，这些约束限制追求目标的程度，所有约束应由线性等式或不等式构成。

·方案选择：必须有两种以上（包括两种）的可行方案供选择。

·决策变量非负：LP问题要求决策变量取值非负（如可行方案中的产品产量不能取负数）。

案例：做一个精明的农场主——图形法

假设一农夫有一块2 000平方米的农地，打算种植小麦或大麦或二者按照某一比例混合种植。该农夫只可以使用有限数量的肥料30 000公斤和农药25 000升，而单位面积的小麦和大麦需要不同数量的肥料和农药。小麦以（F1,P1）（20,10）表示种植单位面积的小麦需要消耗20个单位的肥料和10个单位的农药，大麦以（F2,P2）（30,40）表

示种植单位面积的大麦需要消耗30个单位的肥料和40个单位的农药。设小麦和大麦的销售之后的利润分别为$5和$6，需如何决定种植小麦和大麦的数量以获取最大的利润？

（1）将问题抽象为数学模型

假设小麦和和大麦的种植数量分别是X_1和X_2，以P代表销售大小麦之后的利润，小麦与大麦的种植面积问题用以下线性规划要素来表述。

目标函数（预测变量），即最大化利润$P=5X_1+6X_2(\max)$

约束条件如下。

·限制1：种植面积的限制为$X_1+X_2\leq 2000$，即种植的总面积不能超过可以可种植的总面积。

·限制2：肥料数量的限制为$20X_1+30X_2\leq 30\ 000$，即大小麦消耗的总肥料量不能超过可用的总肥料量。

·限制3：农药数量的限制为$10X_1+40X_2\leq 25\ 000$，即大小麦消耗的总农药量不能超过可用的总农药量。

·限制4：非负性限制为$\{X_1\geq 0且X_2\geq 0\}$

（2）图形法求解最优值

由于限制条件中有$\{X_1\geq 0且X_2\geq 0\}$，所以在绘图过程中X_1和X_2均取正值，即都在坐标轴的第1象限，操作步骤如下。

·限制1：种植面积为$X_1+X_2\leq 2000$，将X_1作为Y轴，X_2作为X轴，则种植面积限制的 表达式转换为$X_1\leq 2000-X_2$，在坐标轴中的表示如图2-44所示。

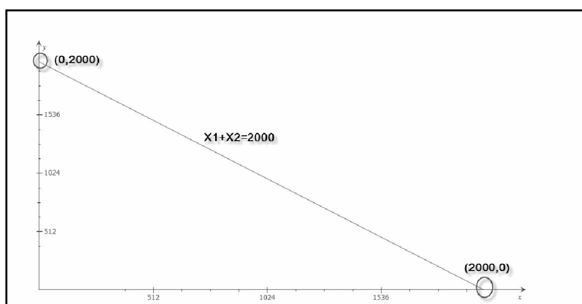


图2-44　种植面积限制

该直线是斜率为-1，在X 轴的截点为（2000,0）；在Y 轴的截点为（0,2000），直线与X 轴和Y 轴中间的任意一点均可以满足该约束条件。

·限制2：肥料数量限制为$20X_1+30X_2\leq 30\ 000$；同样将X_1 作为Y 轴，X_2 作为X 轴，如图2-45所示。

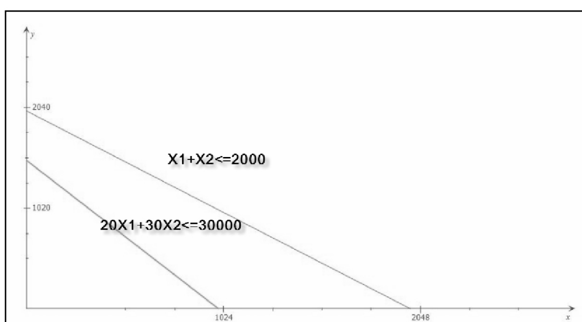


图2-45　两个限制条件2

可以看出限制条件2较限制条件1要更加严格，XY 轴与直线（$20X_1+30X_2\leq 30000$）之间的面积更小。

·限制3：$10X_1+40X_2\leq 25\ 000$；同样将X_1 作为Y 轴，X_2 作为X 轴，如图2-46所示。

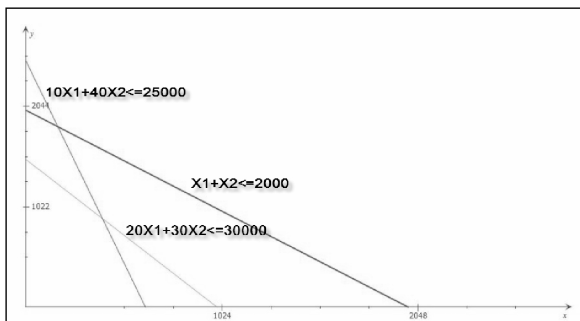


图2-46　3个限制条件

图中所有直线下面积的交集部分即是X_1与X_2的可行区域，如图2-47所示。

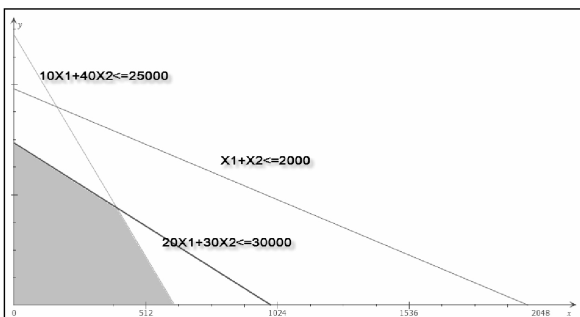


图2-47　可行区域

·转换目标函数

$P=5X_1+6X_2$，将公式变形为$X_1=-\frac{6}{5}X_2+\frac{P}{5}$。将$X_1$作为Y轴，$X_2$作为X轴，则$P/5$为方程在Y轴上的截距。

将$\frac{P}{5}=1000$、$\frac{P}{5}=1500$和$\frac{P}{5}=2000$分别代入方程计算，3条平行线分别与可行区域的相交情况如图2-48所示。

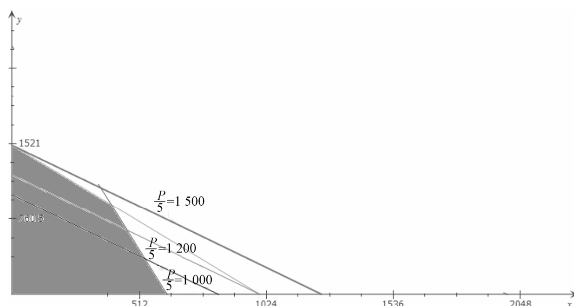


图2-48　等利润法求解极值

当$\frac{P}{5}$=1000、1200和1500时，在图形上是3条平行的等利润线。利润P最大时，$\frac{P}{5}$也最大。从图中可以看出，$\frac{P}{5}$=1500这条平行线在X轴上的截距最大，也与可行区域有交点。因此最大利润点为$\frac{P}{5}$=1500时与Y轴的交点，即$X_2=0$，$X_1=\frac{P}{5}=1500$，利润$P=7500$。

在求解最优解后将$X_1=1500$，$X_2=0$带入各方程，结果如下。

$X_1+X_2=1500\leq 2000$

$20X_1+30X_2=30000\leq 30000$

$10X_1+40X_2=15000\leq 25000$

结果汇总后如表2-3所示。

表**2-3**　结果汇总

约束条件	**$X_1=1500$，$X_2=0$**时需要的各类资源	可用资源数	未用资源
面积	1 500	2 000	500
肥料	30 000	30 000	0

农药	15 000	25 000	10 000

从计算结果可以看出种植面积还有500个单位，农药还有10 000个单位未使用，在此方案 中的“面积”和“农药”这两个变量对“肥料”这个变量是松弛的。在线性规划的术语中a <=约束条件的任何未用产能都称为“与约束条件对应是松弛的”，该变量的松弛部分对目标函数没有贡献。即多余出来的农药和面积对利润没有贡献，此时农场主可以考虑将多余的农药与其他农场主交换多余的肥料，以创造更多的利润。

案例：高尔夫球袋的最优生产计划——图形法

某家公司是一个生产高尔夫器械的小型公司，大部分产品处于低端市场。由于低端市场竞争的加剧，它决定进入中高等价位的高尔夫球袋市场。

市场分销商对其未来的新产品十分感兴趣，且同意买进未来3个月内该公司的全部产品。现在需要生产部安排产品的生产，生产高尔夫球袋主要有4个步骤，即切割并印染原材料、缝合、成型，以及检查和包装。

生产经理详细分析了生产过程的每一步，得出以下结论。

（1）生产一个中价位标准的球袋需要7/10小时切割与印染、1/2小时缝合、1小时成型，以及1/10小时检查和包装。

（2）生产一个高价位球袋需要1小时切割和印染、5/6小时缝合、2/3小时成型，以及1/4小时检查和包装。生产信息汇总如表2-4所示。

表2-4　生产信息汇总

步序	生产时间	
	中级袋	高级袋
切割与印染	7/10	1
缝合	1/2	5/6
成型	1	2/3
检查和包装	1/10	1/4

未来3个月内该公司生产部门各步序的最大生产时间为：切割和印染630小时，缝合600小时，成型708小时，检查与包装135小时。

生产一个标准袋的利润是$10，生产一个高级袋的利润是$9。安排生产各部门的生产计划，以达到公司利润最大的目的操作步骤如下。

（1）转换成数据模型

假设标准袋（中级袋）生产数量为X_1，高级袋的生产数量为X_2，则可将上述问题转化成线性规划问题。

·决策变量：标准袋生产量为X_1，高级袋生产量为X_2。

·目标函数：$Y=10X_1+9X_2$（利润最大化）。

·约束条件：限制条件1为切割和印染总时间≤630小时，即$\frac{7}{10}X_1+X_2\leqslant 630$；限制条件2为缝合总时间≤600小时，即$\frac{7}{2}X_1+\frac{5}{6}X_2\leqslant 600$；限制条件3为成型总时间≤708小时，即$X_1+\frac{2}{3}X_2\leqslant 708$；限制条件4为检查与包装总时间≤135小时，即$\frac{1}{10}X_1+\frac{1}{4}X_2\leqslant 135$；限制条件5为$X_1\geq 0$且$X_2\geq 0$。

（2）图形法求解最优值

利用图形法对上述问题进行求解，以X_1为X轴，以X_2为Y轴。

·限制条件1：切割和印染总时间≤630小时，即 $\frac{7}{10}X_1+X_2\leqslant 630$ ，如图2-49所示。

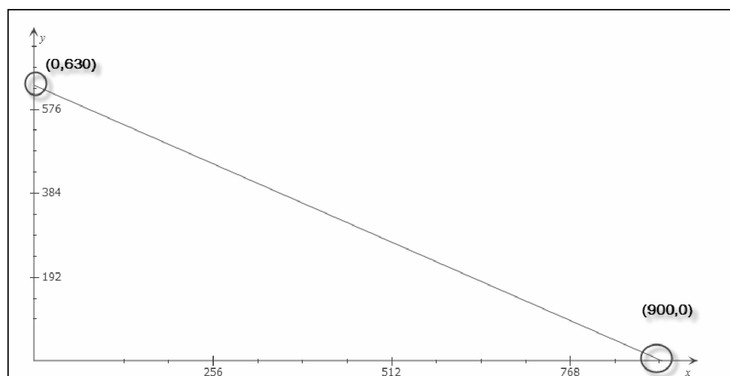


图2-49　切割与印染时间限制

·将其他限制条件一并加入图形中，过程不再一一叙述，在各限制下的公用可行区域如图2-50所示。

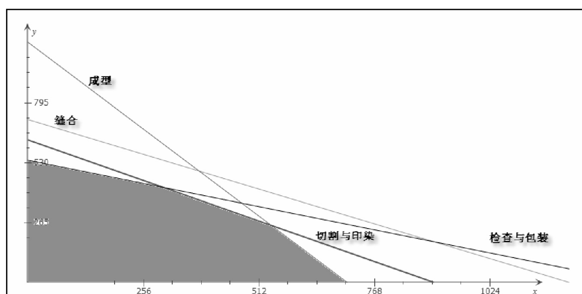


图2-50　各限制条件下的公用可行区域

将不可行区域去除后的可行区域如图2-51所示。

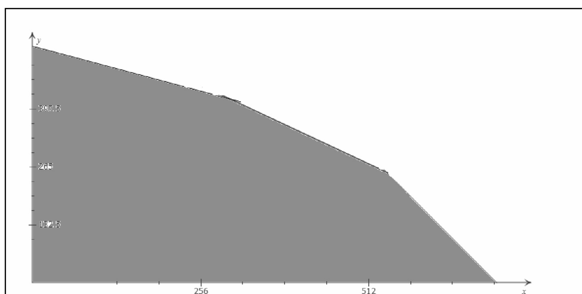


图2-51　去除不可行区域后的可行区域

目标函数$Y=10X_1+9X_2$，转化$X_2=-\frac{10}{9}X_1+\frac{1}{9}Y$，$X_1$作为$X$轴，$X_2$作为$Y$轴，则$Y/9$为方程在$Y$轴上的截距。将$Y=1800$、$Y=3600$和$Y=5400$分别代入方程计算，3条平行的直线与可行区域的相交情况如图2-52所示。

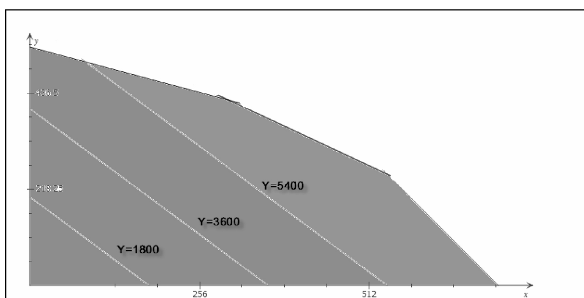


图2-52　3条平行的直线与可行区域的相交情况

从图中可知以上平行的等利润线均不是Y轴截距最大线，将平行线继续向上平移。截距最大处位于平线性和限制条件1与限制条件4的交点处，即同时满足以下方程：

$$\frac{7}{10}X_1+X_2\leqslant 630$$

$$X_1+\frac{2}{3}X_2=708$$

可得X_1 ＝540，X_2 ＝252，则最大利润值为Y ＝10×540＋9×252＝7668，如图2-53所示。

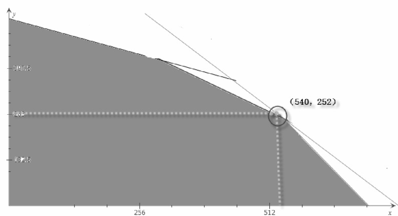


图2-53　最大利润值

（3）理解松弛变量

求解最优解后将X_1 ＝540，X_2 ＝252代入各方程，结果如下。

$$\frac{7}{10}X_1+X_2=630\leqslant 630$$

$$\frac{1}{2}X_1+\frac{5X_2}{6}=480\leqslant 600$$

$$X_1+\frac{2}{3}X_2=700\leqslant 708$$

$$\frac{1}{10}X_1+\frac{1}{4}X_2=117\leqslant 135$$

数据汇总如表2-5所示。

表**2-5**　数据汇总

约束条件	最优解时需要的小时数	可用小时数	未用小时数
切割与印染	630	630	0
缝合	480	600	120
成型	708	708	0
检查与包装	117	135	18

从以上数据可以看出该方案中缝合、检查与包装部门，还有松弛时间，如果管理者需要继续提升产量，则可以考虑将缝合部门的资源

抽调至切割与印染部门或成型部门。这样各部门的资源比较平衡，产能也会发挥到最大。本例中最优值的部门主要是切割与印染和成型部分，如果这两个部门约束条件发生变化，最优值增加量又会有多少？这需要理解敏感性分析和对偶理论。

（4）敏感性分析

当一个线性规划问题的最优解求出以后，如果所给问题的某些条件发生变化，所得的最优解会有什么变化？或者条件变化在什么范围内仍可使所有最优解不变？如果最优解发生了变化，应该如何迅速再次寻求最优解？研究这些问题就是所谓灵敏度分析，即输入X 对输出Y 的最优解的影响程度。由于敏感性分析的目标是输入对最优解的影响，即需要先求出最优解之后的分析，因此敏感性分析也称为“后优化分析”。

在实际生活生产中各个输入变量均在一直不停地变化，而并不是模型中固定的输入变量，因此敏感性分析有助于确定哪些风险对项目具有最大的潜在影响并对决策者至关重要。本例中的利润对输入的多个变量的敏感性如何？下面进行说明。

我们已知道最优解是标准袋生产540个，高级袋生产252个，这个最优解的前提是标准袋的利润是$10且高级袋的利润是$9。如果由于价格变化，标准袋的利润由$10下降至$8.5，这时我们可以用敏感性分析来确定标准袋生产540个和高级袋生产252个是否还是最优解。如果是，则不需要建立新的模型求解。

通常来讲，模型计算出来的数据不可能百分百地接近真实的最优值，只是接近真实值。利润是个估计值，如果高级袋的利润在$6～$14之间比较大的范围波动时模型的最优解都是540个标准袋和252个高级袋，则可以认为模型计算出来的最优解比较可信；如果高级袋的利润在$8.9～$9.2比较窄之间的范围波动时模型的最优解是540个标

准袋和252个高级袋，那么该模型计算出来的最优值的可靠性就值得商榷。

继续用图表法对敏感度进行分析，例中讲到标准袋的利润是$10且高级袋的利润是$9。如果其中一种袋子的利润下降，则应该削减其产量；如果利润上升，则应增加其产量，问题是利润变化多少时我们才改变其产量？

仔细观察最优解的图形，如图2-54所示。

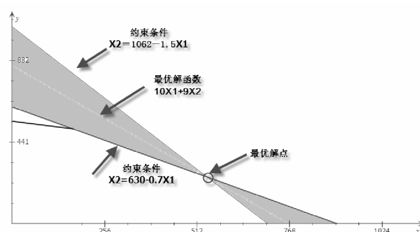


图2-54　最优解的图形

目标解函数是图中的绿色线，该线以最优解为支点上下旋转。只要不超过阴影部分，最优解仍然是图示的最优解点。

逆时针旋转目标函数直线使其斜率变成一个绝对值更小的负数，从而斜率更大。当旋转到足够角度与约束条件$\frac{7}{10}X_1+X_2=630$ 直线重合，即$\frac{7}{10}X_1+X_2=630$ 时该直线的斜率就是目标函数直 线斜率的上限。

同理，将目标函数直线顺时针旋转，使其斜率变成一个绝对值更大的负数，从而斜率更小。当旋转到足够角度与约束条件$X_1+\frac{2}{3}X_2=708$ 直线重合且$X_1+\frac{2}{3}X_2=708$ 时该直线的斜率就是目标函数直线斜率的上限。

$\frac{7}{10}X_1+X_2=630$ 直线的斜率是$\frac{7}{10}$ ，$X_1+\frac{2}{3}X_2=708$ 的斜率是$-\frac{3}{2}$ ，因此目

标函数的斜率的范围为：

$$-\frac{3}{2}<= \text{目标函数的斜率} <=-\frac{7}{10}$$

设标准袋的利润为S，高级袋的利润为D，目标函数的公式为$Y=SX_1+DX_2$，转化$X_2=-\frac{S}{D}X_1+1\frac{Y}{D}$，该目标函数的斜率为$-\frac{S}{D}$。只要-3/2＜＝$-\frac{S}{D}$＜＝-7/10，则方案的最优解仍然不变。

如果高级袋的利润设置为D＝9，则可以推导出标准袋利润S的波动范围为6.3～13.5，计算过程如下：

$-\frac{3}{2}<=-\frac{10}{D}<=-\frac{7}{10}$，则

$6.3\leq S\leq 13.5$

这时，标准袋的利润下降到$6.3或上升到$13.5区间对生产计划的最优值没有影响。

如果标准袋的利润为常数$S=10$，同样可以推导出D的波动区间为6.67～14.29，计算过程如下：

$-\frac{3}{2}<=-\frac{10}{D}<=-\frac{7}{10}$，则

$6.67\leq D\leq 14.29$

这时，高级袋的利润下降到$6.67或上升到$14.29区间对生产计划的最优值没有影响。

如果两种袋子的利润同时在波动，则通过二者比值的负值与上下限的斜率进行比较即可，这样管理着就可以很清楚地了解是否需要调整产品的生产方案。

这就是目标函数总利润对单个产品利润的敏感性分析，即目标函数系数的变化对最优解的影响的分析。

（5）对偶价格

上面只是说明了目标函数的系数的变化对最优解变化的影响，如果约束条件发生变化，对最优解是否又会有什么样的影响？如将上面问题中的切割与印染的约束条件由630调整至640，约束条件变成：

$$\frac{7}{10}X_1+X_2=640$$

其他约束条件不变，可行区域与之前有所变化，如图2-55所示。

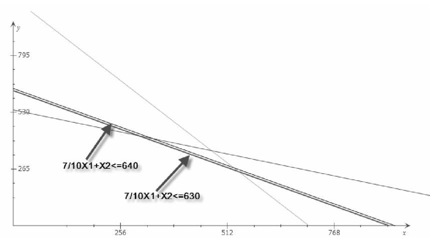


图2-55　约束条件变化后的可行区域

可以看出，增加10小时可行区域较之前有所变宽。重新对最优值进行计算，最优值的点变为 $\frac{7}{10}X_1+X_2=640$ 与 $X_1+\frac{2}{3}X_2=708$ 的交点，求解出为X_1＝527.5，X_2＝270.5。最大利润值为10×527.5＋9×270.5＝7711.75，较之前的7668增加43.75。增加10小时增加的利润为43.75，增加率为43.75/10＝4.375美元／小时。

管理者从分析可以得出的信息是如果将利润最大化，可以将某些松弛变量（缝合部门）的人工时调整到关键的限制条件（切割与印染部门）中。

约束条件的右端值每增加一个单位引起最优值的变化量称为“对偶价格”，如上例中切割与印染的约束条件的对偶价格为4.375美元。切

割与印染的约束条件每增加或减少1个小时，利　润会相应增加或减少4.375美元，当然这只是在一定的范围内，如果该约束条件增加得足够多，即该条件的直线不在可行区域时该约束条件不再对目标函数产生影响，会变成松弛变量。

案例：高尔夫球袋再生产——非线性规划

在上面的求解过程中我们假设生产的标准袋和高级袋均可以全部销售，利润也是固定的。按照正常的市场规律，这个假设是不成立的。销售量与需求量往往是相关的关系，即需求量下降，销售量下降。为了计算利润值，我们需先确定销售量。而销售量又是与价格存在关系。假设X_1是标准袋的销售量，X_2是高级袋的销售量，P_1是标准袋的销售价格，P_2是高级袋的销售价格，销售量与价格之间存在如下关系式：

$$X_1 = 2250\text{-}15P_1$$

$$X_2 = 1500\text{-}15P_2$$

假设每个标准袋的成本是\$70，则生产X_1 个标准袋的成本是$70X_1$；销售X_1 个标准袋的收益为$X_1 P_1$ ，因此生产和销售X_1 个标准袋的利润修正为：

$$X_1 P_1\text{-}70X_1$$

将销售量与价格之间的关系式与利润表达式合并成：

$$X_1P_1-70X_1=(150-\frac{1}{15}X_1)X_1=80X_1-\frac{1}{15}X_1^2$$

同理，假设每个高级袋的生产成本为\$150，高级袋的利润与销售量之间的关系表达式如下所示：

$$X_{2}{}_{1} P_2-150X_2=\left(300-\frac{1}{5}X_2\right)X_2=150X_2-\frac{1}{5}X_2^2$$

总利润为二者的利润相加，得到总利润为：

$$80X_2-\frac{1}{15}X_1^2+150X_2-\frac{1}{5}X_2^2$$

此时该目标函数与销售量之间不再是线性关系，而是非线性关系。将总利润设置为Z轴，X_1 为X轴，X_2 为Y轴，则三者之间的关系如图2-56所示。

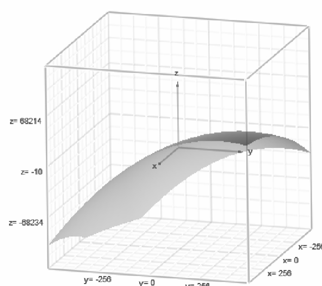


图2-56　利润与销售量关系

在没有约束条件的情况下按照数学方法求解，Z 的最大值是在$X=600$且$Y=375$的位置，即$X_1=600$且$X_2=375$。这点在该案例中的各类约束条件形成的可行解区域，如图2-57所示。

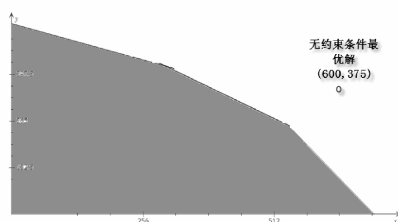


图2-57　无约束条件的解

显然这个最优解不是我们想要的，这是在没有约束条件下使用数

学计算出来的结果，结合前面的约束条件将该问题抽象成如下数学模型：

目标函数：

$$80X_2-\frac{1}{15}{X_1}^2+150X_2-\frac{1}{5}{X_2}^2$$

约束条件：

$$\frac{X_1}{10}+\frac{X_2}{4}\leqslant 135$$

$$\frac{X_1}{2}+\frac{5X_2}{6}\leqslant 600$$

$$X_1+\frac{2X_2}{3}\leqslant 780$$

$$\frac{7X_1}{10}+X_1\leqslant 630$$

此时再利用图形法求解已不太合适，利用LINGO或Crystal Ball可以非常轻松地求解。在线性规划中还有一种较为特殊的规划，即整数线性规划。

案例：做最聪明的房东——整数线性规划

“高尔夫球袋”的例子其实是整数线性规划求解的例子之一，在决策变量中至少有一个变量是整数的这类问题称为“整数线性规划”。如果决策变量中既有整数变量，又有非整数变量，则称为“混合整数线性规划”。如果要求变量等于0或1，这些变量称为“0-1变量”或“二进制变量”，也是整数线性规划的一种；如果所有的变量均为整数；则称之为“全整数线性规划”。

整数线性规划比一般连续变量线性规划求解要难一些，在许多线性规划问题中利用连续变量的求解预测变量在经济性上几乎不产生影

响，所以在不必要情况下尽量不要使用整数线性规划。但是在决定新增几家工厂或生产多少架飞机的情况下近似的影响比较大的时候需要应用整数线性规划方法来解决。

下面通过构建一个含有两个整数变量（X_1 与X_2 ）的全整数线性规划模型来说明：

目标函数：

$$Y = 3X_1 + 2X_2$$

约束条件：

$$3X_1 + 3X_2 \leq 15$$

$$\frac{3}{5}X_1 + 1X_2 \leqslant 8$$

$$2X_1 + 3X_2 \leq 9$$

$$X_1 \text{和} X_2 \geq 0 \text{且为整数}$$

这时就要求决策变量X_1 与X_2 均为整数，该模型称为“整数线性规划模型”。如果将X_1 与X_2 均为整数的约束条件去掉，则该模型变成常规双变量线性规划；如果只有一个变量是整数，则变成混合整数线性规划模型。

目标函数：

$$Y = 3X_1 + 2X_2$$

约束条件：

$$3X_1 + 3X_2 \leq 15$$

$$\frac{3}{5}X_1+1X_2\leq 8$$

$$2X_1+3X_2\leq 9$$

X_1和X_2均≥ 0且为整数。

下面以应用实例来说明全整数线性规划的求解。

某房地产公司有$200万用于购买新的租赁财产，经过内部筛选公司已将投资项目的方案减少为联体别墅和公寓楼。每套联体别墅的售价为$282　　000，现市面上有5套空置的别墅。每栋公寓楼的售价为$400 000，而且开发商可以根据房产公司的需求数量再建造房子。

房产公司每月有140小时处理新购置的房产，其中每套联体别墅预计每月需花费4小时，每栋公寓楼每月40小时。在扣除抵押偿还和经营成本后，房产公司出租这些房产将租金预计为每套联体别墅$10 000且每栋公寓楼$15　000。房产公司股东想知道在保证租金最大的要求下，购买联体别墅和公寓楼的数量分别应该为多少？

将问题抽象成数学模型，决策变量X_1为连体别墅的购买数量，X_2为公寓楼的购买数量。

预测变量（目标函数）为每年的租金总量（单位：$1 000）：

$$Y=10X_1+15X_2$$

该函数的4个约束条件是：

$282X_1+400X_2\leq 2000$（总资金量）

$4X_1+40X_2\leq 140$（每月的管理时间）

$X_1\leq 5$（总联体别墅数）

X_1 和$X_2 \geq 0$且X_1 为整数（非负与整数要求）。

我们现在去掉决策变量为整数的要求，暂时按照连续变量的线性规划求解方式计算。按照前面介绍的图形法求解步骤（求解过程不再一一叙述）得到的最优解为$X_1 = 2.479$套联体别墅，$X_2 = 3.252$套公寓楼。Y 的最大值为73.574，图解分析结果如图2-58所示。

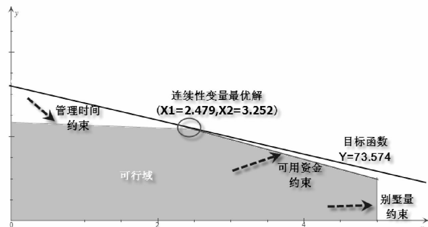


图2-58　图解分析结果

在这种情况下我们一般的处理方式是按照近似取整数的方式进行的，但在决策变量微小变化对预测变量产生较大影响时不适合采取这种处理方式，上例是典型的这种情况。别墅和公寓楼的销售价格均比较高，而能购买的数量又相对较少。如果近似得到一个整数解，将会对预测变量总租金产生较大的影响。

假设我们把最优解近似为$X_1 = 2$且$X_2 = 3$，于是目标函数Y 的值变为$10 \times 2 + 15 \times 3 = 65$，与连续性求解情况下的73.754相差较大。将其他近似整数值与Y 值和连续性线性规划的数值进行对比，如表2-6所示。

表**2-6**　最优解对比

X_1 值	X_2 值	Y 值	说　明
2.479	3.252	73.574	连续性最优解
2	3	65	与最优解相差较大
3	3	75	购房款需要204

			万美元，超出房产公司购房总资金
2	4	80	购房款需要216万美元，超出房产公司购房总资金

从以上结果看好像只有$X_1=2$和$X_2=3$是可以满足约束条件的最优解，那么该方案是不是最优？我们继续按照图解法对整数线性规划进行计算。由于此时限定了决策变量是整数型，因此Y值也一定是整数型，求解结果如图2-59所示。

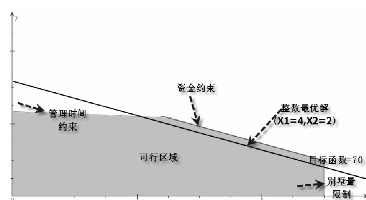


图2-59　整数型最优解结果

从图中可以看出当$X_1=4$且$X_2=2$时得到最优整数值，目标函数值为$10\times4+15\times2=70$，即年总租金为$70 000。这一结果要比近似得到的$X_1=2$且$X_2=3$时总租金为$65 000好很多，因此此类问题需要通过整数线性规划求解来计算最优值。

与非线性规划一样，如果遇到整数线性规划则使用图形法不能方便地得出结果，此时需要使用软件。

Mr Shu：“怎么样？现在对线性规划求解是不是有了一个比较

深刻的认识？”

Miss Ju：“通过图形求解还是能比较清晰地看清楚过程。”

Mr Shu：“你发现没有线性规划的求解其实一点也不难，很好理解。”

Miss Ju：“是的，优化求解都是基于这个原理吗？”

Mr Shu：“线性规划基于这些原理，如果存在非线性规划或整数线性规划，则使用图形求解比较困难。在上面的例子中我们也看到了，但很多都是在非线性的情况下求解最优解。”

Miss Ju：“非线性时我们应该如何来求解？”

Mr Shu：“不要担心，我们还有Crystal Ball和LINGO。之前我们已经了解过CB软件在风险和不确定性分析方面的一些应用，下面我们来看看它在优化求解方面的巨大威力。还有另一款在优化求解方面大名鼎鼎的LINGO，看看它可以给我们带来什么惊喜吧。”

2.3　见识LINGO与Crystal Ball的威力

LINGO（Linear Interactive and General Optimizer，交互式的线性和通用优化求解器）由美国LINDO系统公司（Lindo System Inc.）推出，它可以用于求解非线性规划，以及一些线性和非线性方程组的求解等。其功能十分强大，是求解优化模型的绝佳选择。

LINGO的特色在于内置建模语言，提供的十几个内部函数允许决

策变量是整数（即整数规划，包括0-1整数规划），方便灵活且执行速度非常快。它能方便地与Excel和数据库等其他软件交换数据，易于方便地输入、求解和分析大规模最优化问题。

由于这些特点，LINDO系统公司的线性、非线性和整数规划求解程序已经被全世界数千万的公司用来分析最大化利润和最小化成本，其应用的范围包含生产线规划、运输、财务金融、投资分配、资本预算、混合排程、库存管理和资源配置等。

一般来讲，使用LINGO求解运筹学问题可以分为以下两个步骤完成。

（1）根据实际问题建立数学模型，即使用数学建模的方法建立优化模型。

（2）根据优化模型利用LINGO来求解模型，主要是根据LINGO软件把数学模型转译成计算机语言，借助计算机来求解。

下面通过几个案例来看看CB和LINGO如何完成这些复杂任务。

案例：大麦种植新解——LINGO求解

继续以农夫种田的案例为例，在LINGO中的操作步骤如下。

（1）整理数学模型

目标函数（预测变量）为最大化利润$P=5X_1+6X_2\,(\max)$。

约束条件如下。

·限制1：种植面积的限制为$X_1+X_2\leq 2000$。

·限制2：肥料数量的限制为$20X_1+30X_2\leq 30000$。

·限制3：农药数量的限制为$10X_1 + 40X_2 \leq 25000$。

·限制4：非负性限制为$\{X_1 \geq 0且X_2 \geq 0\}$。

（2）输入模型

将以下方程式输入至LINGO软件中：

$$MAX = 5*X_1 + 6*X_2 ;$$

$$X_1 + X_2 <= 2000 ;$$

$$20*X_1 + 30*X_2 <= 30000 ;$$

$$10*X_1 + 40*X_2 <= 25000 ;$$

打开LINGO软件，在【Model】中输入以上方程式，如图2-60所示。



图2-60　输入方程式

（3）求解

单击LINGO工具栏中的【Solve】按钮，计算结果如图2-61所示。

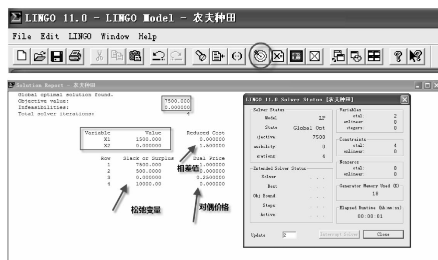


图2-61　LINGO计算结果

从以上计算结果可以看出与图形法求解的结果一样；同时给出了一些更加丰富的信息，如松弛变量、对偶价格和相差值等。

相差值类似前面说的敏感性分析，X_2 求解出来的值为0，相差值为1.5。即如果大麦的利润能增加1.5到7.5，就会对总利润产生影响，可以种植。

利用LINGO软件我们可以非常轻松地求解出结果，在使用该软件时应注意如下事项：

（1）LINGO总是根据“MAX＝”或“MIN＝”寻找目标函数，而除注释语句和TITLE语句外的其他语句都是约束条件，因此语句的顺序并不重要。

（2）LINGO中的函数一律需要以“@”开头，其中整型变量函数（@BIN和@GIN）和上下界限定函数（@FREE、@SUB和@SLB）与LINDO中的命令类似，而且0/1变量函数是@BIN函数。

（3）LINGO中不区分大小写字母；变量和行名可以超过8个字符，但不能超过32个字符且必须以字母开头。

（4）变量可以放在约束条件的右端（数字也可放在约束条件的左端），但为了提高LINGO求解时的效率，如果可能，应尽可能采用

线性表达式定义目标和约束。

（5）语句是组成LINGO模型的基本单位，每个语句都以分号结尾，编写程序时应注意模型的可读性。例如，一行只写一个语句并按照语句之间的嵌套关系适当缩进语句，增强层次感。

（6）以感叹号开始的是说明语句，也需要以分号结束。

LINGO的入门操作很直观也很简单，但是这些还不能真正发挥其威力，它最强大的功能在于使用集合和属性，可以将大型的数学模型简单化。本书作为简单的数据分析入门级，只简单介绍其集合和属性功能，重点介绍一些函数。

LINGO的主要操作界面如图2-62所示。



图2-62　LINGO的主要操作界面

案例：大麦种植新解——**Crystal Ball**求解

在了解使用图形法和LINGO软件求解农夫种田的案例后，使用Crystal Ball同样可以计算。

在第1章中我们了解到在Crystal Ball软件建模过程中有3种变量，即假设变量、决策变量和预测变量（目标变量）需要设置。假设变量通常定义为随机变量，即不随人的意志变化。在 本案例中没有假设变量，因此只需要设置决策变量和预测变量即可。预测变量即目标值，如利润、产值和成本等；决策变量是在计算过程中我们可以控制并且对预测变量产生重大影响的变量，如生产开工率等。操作步骤如下。

（1）设置决策变量

新建一个Excel工作表，选择任一单元格，本例选择A4单元格为决策变量中的小麦数量的求解。在Excel菜单中单击【Crystal Ball】|【Define Decision】选项，显示【Define Decision】对话框，如图2-63所示。



图2-63 【Define Decision】对话框

在【Name】下拉列表框中输入“小麦数量”，在【Bounds】选项组中设置决策变量的求解范围。目前我们不知道应该范围，因此设置为Lower为0，Upper为2000。

执行同样操作步骤设置“大麦数量”，如图2-64所示。

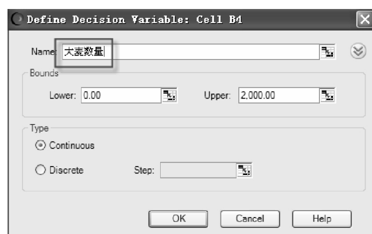


图2-64 设置“大麦数量”

设置这两个决策变量的数据类型均为Continuous变量，即连续类型。

（2）设置目标函数和限制条件

选择A5单元格，输入公式“=5*A4+6*B4”。

在Excel菜单中选择【Crystal Ball】选项，显示【Define Forecast】对话框，如图2-65所示。

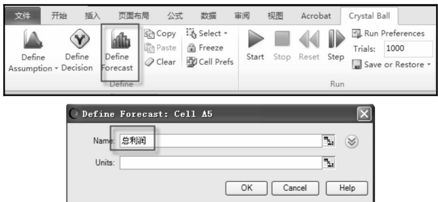


图2-65　【Define Forecast】对话框

设置目标函数的名称为“总利润”。在B6:B8单元格设置限制条件，B6设置公式“=A4+B4”，B7设置公式“=20*A4+30*B4”，B8设置公式“=10*A4+40*B4”。

（3）优化计算

选择【Crystal Ball】|【OptQuest】选项，设置有关选项，如图2-66所示。

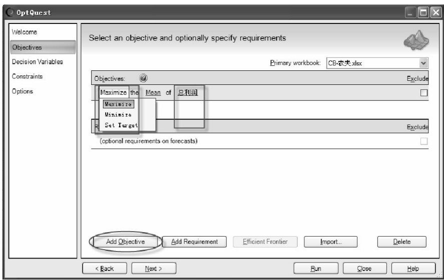


图2-66　设置有关选项

单击【Add Objective】按钮添加目标，选择Maximize（最大化）目标函数“总利润”。单击【Next】按钮，查看决策变量设置，如

图2-67所示。

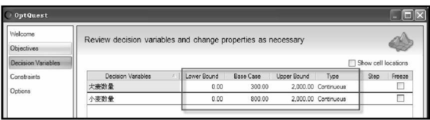


图2-67　查看决策变量设置

以上信息用于说明决策变量的设置范围、变量初值和数据类型。【Freeze】按钮在需要屏蔽某些决策变量时使用，选择即可屏蔽，本例中不选择。单击【Next】按钮，显示【OptQuest】窗口，如图2-68所示。

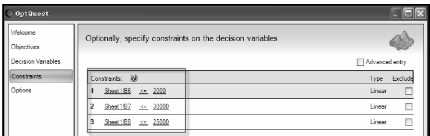


图2-68　【OptQuest】窗口

该窗口用于设置约束条件，选择B6单元格≤2000，表示总的种植面积不能超过2　000；另外两个约束条件依次类推。单击【Next】按钮，单击【Run】按钮进行优化计算，结果如图2-69所示。

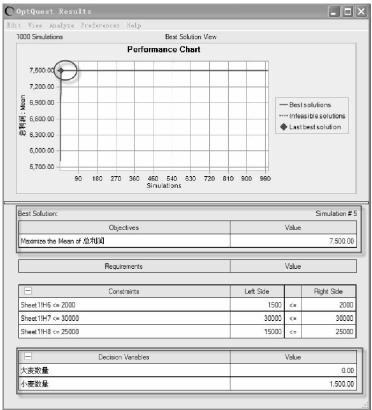


图2-69　优化计算结果

从计算结果来看，Crystal　Ball与LINGO，以及图形法一致。利用软件计算则特别方便。再来看看在这两款软件在非线性优化计算中的应用。

案例：高尔夫球袋新解——LINGO非线性规划

根据2.2节中所述，高尔夫球袋的利润模型修正为非线性之后的为：

$$80X_1-\frac{1}{15}X_1^2+150X_2-\frac{1}{5}X_2^2$$

在LINGO中求解最佳值的操作步骤如下。

（1）输入模型

在LINGO中输入如下内容。

max＝80*x1-x1*x1/15＋150*X2-0.2*X2*X2；

0.1*X1＋0.25*x2＜＝135；

0.5*x1＋5*X2/6＜＝600；

x1＋2*X2/3＜＝708；

0.7*x1＋X2＜＝630；

End

如图2-70所示。

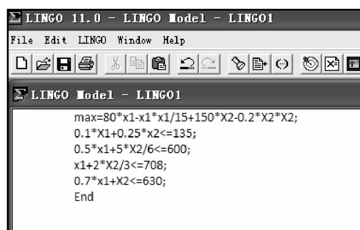


图2-70　输入模型

（2）求解最优值

单击LINGO工具栏中的【Solve】按钮，计算结果如图2-71所示。

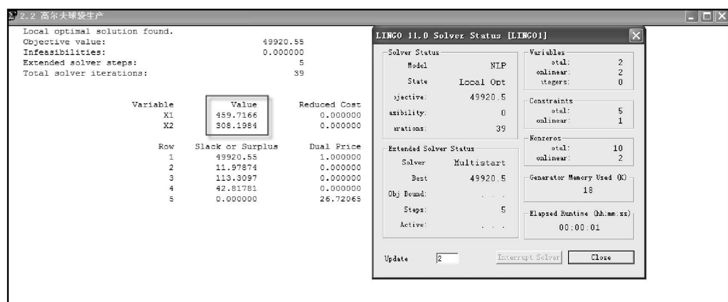


图2-71　计算结果

从计算结果来看标准袋需要生产459.7个；高级袋需要生产308.2个，总利润最大值为49920。当然我们可以对球袋的个数进行整数限制。

案例：一起来倒卖原油——LINGO求解

某公司用两种原油（A和B）混合加工成两种汽油（甲和乙），甲和乙两种汽油含原油A的最低比例分别为50%和60%，每吨售价分别为4 800元和5 600元。该公司现有原油A和B的库存量分别为500吨和

1 000吨，还可以从市场上买到不超过1 500吨的原油A。原油A的市场价为购买量不超过500吨时的单价为10 000元／吨；购买量超过500吨，但不超过1 000吨时超过500吨的部分为8 000元／吨；购买量超过1 000吨时，超过1 000吨的部分为6 000元／吨，该公司应如何安排原油的采购和加工？

这个案例与我们之前用Excel规划求解的原油混合的案例非常类似，不同的是采购价格不再按照单价计算。而是分段函数，属于非线性规划。在利用软件求解之前，先对案例进行分析。

安排原油采购和加工的目标是利润最大化，案例中给出的是两种汽油的售价和原油A的采购价，利润为销售汽油的收入与购买原油A的支出之差。这里的难点在于原油A的采购价与购买量的关系比较复杂，是分段函数关系，导致目标函数总利润也是非线性。

建立模型设原油A的购买量为x（吨），根据题目所给数据采购支出$c(x)$可表示为如下分段线性函数（以下价格以千元／吨为单位）：

$$C(\chi)=\begin{cases}10\chi & 0\leqslant\chi\leqslant 500\\ 1000+8\chi & 500\leqslant\chi\leqslant 1000\\ 3000+6\chi & 1000\leqslant\chi\leqslant 1500\end{cases}$$

设原油A用于生产甲和乙两种汽油的数量分别为x_{11}和x_{12}（吨）且原油B用于生产甲和乙两种汽油的数量分别为χ_{21}和χ_{22}（吨），则总收入为$4.8(\chi_{11}+\chi_{21})+5.6(\chi_{12}+\chi_{22})$（千元），于是本例的目标函数（利润）为：

$$\text{Max } 4.8(\chi_{11}+\chi_{21})+5.6(\chi_{12}+\chi_{22})-c(\chi)$$

约束条件包括加工两种汽油用的原油A和原油B库存量的限制、原油A购买量的限制，以及两种汽油含原油A的比例限制，它们表示为：

$$\chi_{11}+\chi_{12}\leqslant\chi+500$$

$$\chi_{21}+\chi_{22}\leqslant 1000$$

$$\chi\leqslant 1500$$

$$\frac{\chi_{11}}{\chi_{11}+\chi_{12}}\geqslant 0.5$$

$$\frac{\chi_{12}}{\chi_{12}+\chi_{22}}\geqslant 0.6$$

$$\chi,\chi_{11},\chi_{12},\chi_{21},\chi_{22}\geqslant 0$$

这样用分段函数定义的$c(\chi)$一般的非线性规划软件难以输入和求解，是否设法将该模型化简，从而用软件求解？

在LINGO软件中的操作步骤如下。

（1）转换数学模型

将原油A的采购量x分解为3个量，即用x_1、x_2和x_3分别表示以价格10、8，以及6千元/吨采购的原油A的吨数，总支出为$c(x)=10x_1+8x_2+6x_3$，且：

$$\chi=\chi_1+\chi_2+\chi_3$$

目标函数变为线性函数：

$$\text{Max } 4.8(\chi_{11}+\chi_{22})+5.6(\chi_{12}+\chi_{22})-(10\chi_1+8\chi_2+6\chi_3)$$

应该注意到只有当以10千元/吨的价格购买$x_1=500$（吨）时，才能以8千元/吨的价格购买$x_2(>0)$，这个限制条件可以表示为：

$$(\chi_1-500)\chi_2=0$$

同理，只有当以8千元/吨的价格购买$x_2=500$（吨）时才能以6千元/吨的价格购买$X_3(>0)$，于是：

$$(\chi_2-500)\chi_3=0$$

此外，x_1、x_2和x_3的取值范围是：

$$0\leq\chi_1,\chi_2,\chi_3,\leq 500$$

因此该模型就有7个决策变量、1目标函数和7个约束条件。在LINGO软件中输入以下模型：

$$Max=4.8*x_{11}+4.8*x_{21}+5.6*x_{12}+5.6*x_{22}-10*x_{1}-8*x_{2}-6*x_{3};$$

$$x_{11}+x_{12}<x+500;$$

$$x_{21}+x_{22}<1000;$$

$$0.5*x11-0.5*x21>0;$$

$$0.4*x_{12}-0.6*x_{22}>0;$$

$$x=x_{1}+x_{2}+x_{3};$$

$$(x_{1}-500)*x_{2}=0;$$

$$(x_{2}-500)*x_{3}=0;$$

$$@bnd(0,x_{1},500);$$

$$@bnd(0,x_{2},500);$$

$$@bnd(0,x_{3},500);$$

End

（2）输入数学模型

将以上模型输入至LINGO软件中，如图2-72所示。

@bnd是LINGO的定义变量范围的函数，@bnd(0,x_{1},500）表示x 1的取值范围为0～500，x_{2}与x_{3}设置与x_{1}相同。

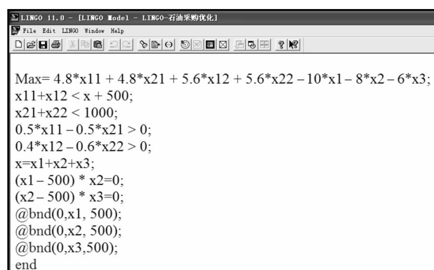


图2-72　输入数学模型

（3）求解最优值

单击【Solve】按钮，计算结果如图2-73所示。

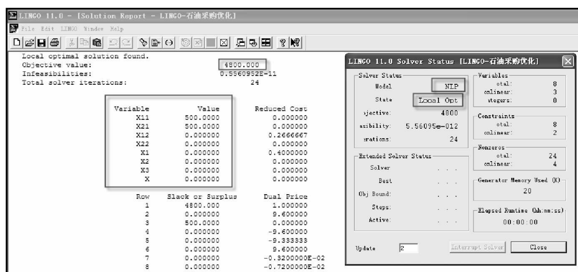


图2-73　最优值计算结果

最优解为用库存的500吨原油A和500吨原油B生产1 000吨汽油甲且不购买新的原油A，利润为4 800（千元）。

但是我们注意在计算结果面板中的【State】文本框中显示“Local Opt”，即局部最优解，而我们需要的全局最优解。

（4）求解全局最优值

选择【LINGO】|【Options】选项，显示【LINGO Options】对话框。选择【Global Solver】选项卡中启动全局优化的【Use Global Solver】复选框，如图2-74所示。

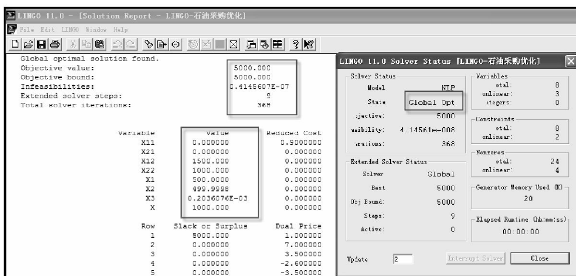


图2-74　设置全局最优解

选择【LINGO】|【Solve】选项，计算结果如图2-75所示。

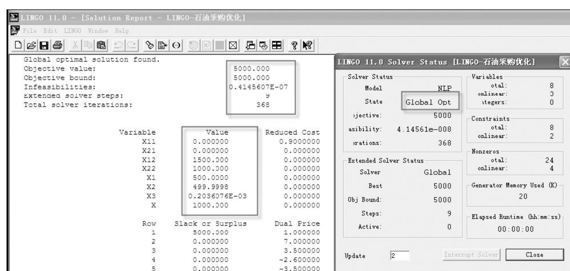


图2-75　全局最优解计算结果

此时LINGO得到的结果是一个全局最优解（Global optimal solution），即购买1 000吨原油A，与库存的500吨原油A和1 000吨原油B一起共生产2 500吨汽油乙，利润为5 000（千元），高于刚刚得到的局部最优解对应的利润4 800（千元）。

案例：一起来倒卖原油——Crystal Ball求解

（1）建立模型

使用LINGO软件求解时我们设置了7个决策变量，在Crystal Ball软件中我们设置5个决策变量，即原油A采购量、A调配到甲数量、A调配到乙数量、B调配到甲数量和B调配到乙数量，Excel中的数据模

型如图2-76所示。

采购设置		甲产品调和				
原油A采购量	原油A初始库存	A到甲	B到甲	原料A成分	限制	产品数量
800	500	150	200	43%	50%	350
采购量限制	原油B初始库存	乙产品调和				
1500	1000	A到乙	B到乙	原料A成分	限制	产品数量
原油A采购支出	7400	150	200	43%	60%	350
销售产值	3640	原料A总消耗	原料B总消耗			
总利润	-3760	300	400			
原料A总容量	1300					

图2-76　Crystal Ball的求解原油采购模型

模型建立在“石油采购优化.xls”工作表中，建立过程略，对模型说明如下。

·B4单元格设置为决策变量，原油A采购量的初始值为800。

·C4单元格设置为原油A的原始库存500吨。

·B6单元格设置为采购限制量1 500吨。

·C6单元格设置为原油B的初始库存1 000吨。

·D4单元格设置为决策变量原油A调配到甲数量，初始值为150。

·E4单元格设置为决策变量原油B调配到甲数量，初始值为200。

·F4单元格设置为甲成品油中A组分含量，在其中输入公式“＝(D4/(D4＋E4))”。

·G4单元格设置甲成品油种A组分含量限制条件50%。

·H4单元格设置甲成品油数量，输入公式“＝D4＋E4”。

·D7单元格设置为决策变量原油A调配到乙数量，初始值为150。

·E7单元格设置为决策变量原油B调配到乙数量，初始值为200。

·F4单元格设置为乙成品油中A组分含量，输入公式“＝D7/(D7＋E7)”。

·G7单元格设置乙成品油种A组分含量限制条件60%。

·H7单元格设置乙成品油数量，在其中输入公式“＝D7＋E7”。

·C7单元格设置成原油A采购支出，在其中输入公式

“＝IF(B4＜＝500,B4*10,IF(B4＜＝1000,1000＋8*B4,IF(B4＜＝1500,30

即采购支出的分段函数，通过Excel的IF函数可以轻松完成。

·C8单元格设置为销售产值，在其中输入公式“＝4.8*H4＋5.6*H7”。

·D9单元格设置为原油A的总消耗，在其中输入公式“＝D4＋D7”。

·E9单元格设置成原油B的总消耗，在其中输入公式“＝E4＋E7”。

·C11单元格作为约束条件，在其中输入公式“＝B4＋C4”。

·C9单元格作为模型的目标函数，在其中输入公式“＝C8-C7”。

（2）设置决策变量

设置决策变量，如图2-77所示。

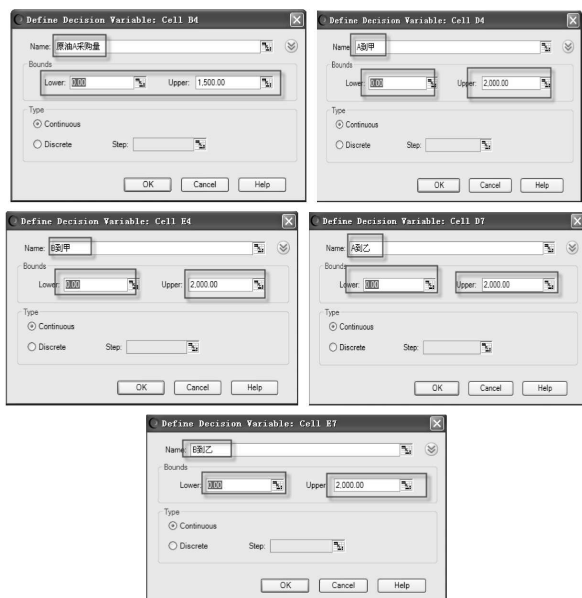


图2-77　设置决策变量

原油A的采购量由于最多只能购买1500吨，因此在设置决策变量时设置为Lower为0且Upper为2000，其他4个决策变量的范围均设置为0～2000。

设置预测变量（目标函数），如图2-78所示。

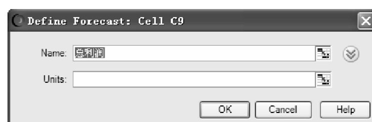


图2-78　设置目标函数

（3）优化计算

在Excel菜单中单击【Crystal　Ball】|【OptQuest】选项，设置及优化计算结果如图2-79所示。

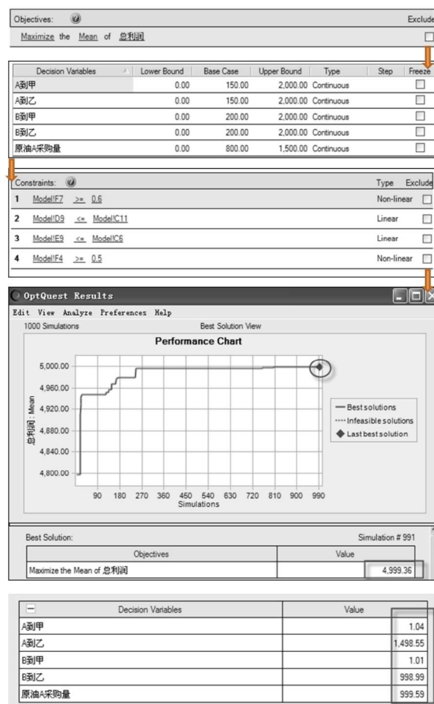


图2-79　设置及优化计算结果

Crystal Ball（CB）计算的总利润的结果为4999.36。

从结果看两款软件计算有细微的差别，即LINGO计算出总利润最大值为5 000，原油A的购买量为1 000；CB的计算结果为4 999.36，原油A的购买量为999.59。究其原因，是我们在CB中设置决策变量为“Continous”（连续性）。如果我们设置成“Discrete”（离散型）计算，则结果会保持一致。

Miss Ju：“看来LINGO与Crystal Ball在求解最优值方面各有优势，前者的输入比较简便，界面简单，难点在于梳理数学模型；后者可以结合Excel的强大函数功能，操作比较灵活。”

Mr　Shu：“是的，不管使用哪种软件都需要整理数学模型。LINGO的强大之处还没有体现出来，它可以处理一些非常大型的数学模型，但要求我们有属性和集合的概念。遇到这样的问题时，如果使用Crystal　Ball处理，一般会需要设立很多决策变量，而使用LINGO则可以非常简便和清晰地整理出来。”

Miss　Ju：“二者的特点的确不同，能不能列举使用LINGO关于属性和集合简单的例子，再看看Crystal　Ball在优化求解方面的特点？”

Mr Shu：“当然可以，Miss Ju，有没有扬帆远航的想法？”

Miss Ju：“一直有，难道LINGO还可以让我们出海远航吗？”

Mr Shu：“当然不是，只是我们来借用一个帆船的案例来见识一下LINGO在处理较大型模型的方法，再来看看Crystal　Ball还有哪些惊喜可以带给我们。”

案例：扬帆远航——帆船生产计划优化（LINGO求解）

每到年底SAILCO公司就需要决定下下一年，即下4个季度的帆船生产量。从市场反馈的情况来看，下4个季度的帆船需求量分别是40艘、60艘、75艘和25艘，这些需求必须按时满足。

按照SAILCO公司的生产能力，每个季度正常的生产能力是40艘帆船，每艘船的生产费用为400美元。如果加班生产，每艘船的生产费用为450美元，每个季度末每艘船的库存费用为20美元。假定生产提前期为0，初始库存为10艘船。如何安排生产计划可使生产总费用最小？

这是个典型的优化计算，使用Crystal　Ball也可以非常迅速地求解

出结果，但是我们这里运用LINGO的集合和属性来解答。

（1）模型整理

分别用英文的头几个字母表示正常的需求量（DEM）、正常生产的产量（RP）、加班生产的产量（OP）和库存量（INV），则每个季度都对应有相应的数值。即它们都应该是一个由4个元素组成的数组，其中DEM是已知的，而RP、OP和INV是未知数。

目标函数是所有费用的和，即：

$$400PR(I)+450OP(I)+20INV(I)$$

I分别表示1～4共4个季度。

约束条件如下。

·生产能力限制：$PR(I)<40$且$I=1,2,3,4$。

·产品数量的平衡：$INV(I)=INV(I-1)+PR(I)+OP(I)-DEM(I)$且$I=1,2,3,4$。

另外加上各变量的非负限制。

LINDO中没有数组，只能为每个季度分别定义变量，如正常产量要有RP 1、RP 2、RP 3和RP 4共4个变量等。写起来比较麻烦，尤其是更多（如1 000个季度）时。

记4个季度组成的集合$TIMES=\{1,2,3,4\}$，它们是上面数组的下标集合，而数组DEM、RP、OP和INV对应集合$TIMES$中的每个元素1～4分别有一个值。LINGO正是充分利用了这种数组及其下标的关系引入了“集合”及其“属性”的概念，把TIMES＝{1,2,3,4}称为“集

合”，把*DEM* 、*RP* 、*OP* 和*INV* 称为“该集合的属性”（即定义在该集合上的属性）。可以简单把属性理解为集合中的相同特性，如性别和身高等，而本例中的每个季度都有4个属性。

集合与属性之间的关联如表2-7所示。

表2-7　集合与属性之间的关联

集合 TIMES 的元素		1	2	3	4
定义在集合 *TIMES* 上的属性	*DEM*	*DEM*（1）	*DEM*（2）	*DEM*（3）	*DEM*（4）
	RP	*RP*（1）	*RP*（2）	*RP*（3）	*RP*（4）
	OP	*OP*（1）	*OP*（2）	*OP*（3）	*OP*（4）
	INV	*INV*（1）	*INV*（2）	*INV*（3）	*INV*（4）

整理的数据模型如下。

```
Model：
sets：
TIMES/1，2，3，4/：DEM，PR，OP，INV；
ENDSETS
MIN=@SUM（TIMES：400*PR+450*OP+20*INV）；
@FOR（TIMES（I）：PR（I）<40）；
@FOR（TIMES（I）|I#GT#1：
INV（I）=INV（I-1）+PR（I）+OP（I）-DEM（I）；
INV（1）=10+PR（1）+OP（1）-DEM（1）；）；
DATA：
DEM=40，60，75，25；
ENDDATA
END
```

有了集合和属性之后LINGO定义模型与普通模型定义不一样，在定义集合和属性有如下注意事项。

·模型以“Model”开头，集合和属性定义部分从“Sets”开始到“Endsets”结束。

·初始值以“DATA”开头，以“ENDDATA”结束，如本例中的市场需求量。

·目标函数的定义方式为@SUM（集合（下标）：关于集合的属性表达式），对语句中冒号“：”后面的表达式按照“：”前面的集合指定的下标（元素）求和。本例中目标函数也可以等价写成@SUM(*TIMES* (*i*)：400***RP* (*i*)+450***OP* (*i*)+20***INV* (*i*))，“@SUM”相当于求和符号“Σ”。目标函数对集合QUARTERS中的所有元素（下标）都要求和，所以可以将下标*i* 省去。

·约束条件定义方式为循环函数@FOR（集合（下标）：关于集合的属性的约束关系式）。本例中，每个季度正常的生产能力是40艘帆船，这正是语句“@FOR(*TIMES* (*I*)：*RP* (*I*)<40)；”的含义。由于对所有元素（下标*I* ），约束的形式是一样的，所以也可以像上面定义目标函数时一样将下标*i* 省去，这个语句可以简化成“@FOR(*TIMES* ：*RP* <40)”。

·本例中对于产品数量的平衡方程，由于下标*i* =1时的约束关系式与*i* =2、3或4时有所区别，所以不能省略下标“*i* ”。实际上，*i* =1时要用到变量*INV* (0)。但定义的属性变量中*INV* 不包含*INV* (0)(*INV* (0)=10，它是一个已知的）。

·为了区别*i* =1和*i* =2、3或4，把*i* =1时的约束关系式单独写出，即“*INV* (1)=10+*RP* (1)+*OP* (1)-*DEM* (1)；”。而对*i* =2、3或4对应的约束，对下标集合的元素（下标*i*)增加了一个逻辑关系式“*i* #GT#1”（这个限制条件与集合之间用一个竖

线“|”分开，称为“过滤条件”）。它是一个逻辑表达式，表示$i>1$；“#GT#”是逻辑运算符号，表示“大于”（Greater Than的字首字母缩写）。

（2）LINGO求解

将以上数学模型输入LINGO中，单击【Solve】按钮并求解，结果如图2-80所示。

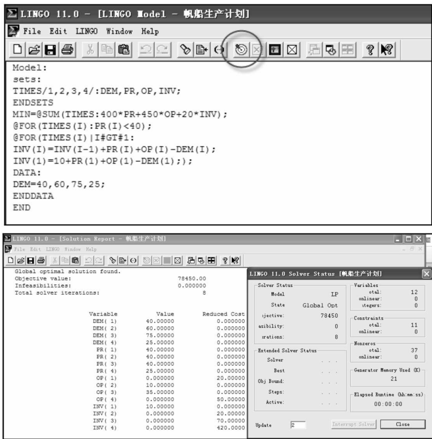


图2-80 帆船计划优化结算结果

从计算结果来看最低成本为78450，各决策变量的结算结果如表2-8所示。

表**2-8** 各决策变量的结算结果

变量	第1季度	第2季度	第3季度	第4季度
DEM	40	60	75	25
RP	40	40	40	25
OP	0	10	35	0
INV	10	0	0	0

本例中共有16个决策变量，在LINGO中我们只定义了一个集合和4个属性就完成了16个决策变量的计算。如果不是4个季度，而是100个季度，此时的决策变量就会变为400个。而使用LINGO软件求解时同样只需要一个集合和4个属性，这就是该软件的强大之处。

一般来说，在LINGO中建立的优化模型可以由5个部分组成或称为“5段”（SECTION）。

（1）集合段（SETS）：以“SETS：”开始，以“ENDSETS”结束，其中定义必要的集合变量（SET）及其元素（MEMBER，含义类似数组下标）和属性（ATTRIBUTE，含义类似数组）。

如上例中定义了集合*TIMES*（含义是季节），它包含4个元素，即4个周期指标（1、2、3和4），每个周期都有需求（*DEM*）、正常生产量（*RP*）、加班生产量（*OP*）和库存量（*INV*）等属性（相当于数组，数组下标由TIMES元素决定）。一旦建立这样的定义，如果*TIMES*的数量不是4，而是1 000，只需扩展其元素为1，2，…，1000，每个周期仍然都有*DEM*、*RP*、OP和*INV*这样的属性（如果这些量定常量，则可在数据段输入；如果是未知数，则可在初始段输入初值）。当*TIMES*的数量不是4，而是1 000时，没有必要把1，2，…，1000逐个列出，而是如下定义TIMES集合。即“*TIMES* /1..1000/:*DEM*,*RP*,*OP*,*INV*;”，“1..1000”表示1～1 000的所有整数。

（2）目标与约束段：目标函数和约束条件等，没有段的开始和结束标记，因此实际上是除其他4个段（都有明确的段标记）外的LINGO模型。

这里一般要用到LINGO的内部函数，尤其是与集合相关的求和函数@SUM和循环函数@FOR等。

上例中定义的目标函数与quarters的元素数目是4或1　000并无具体的关系，约束的表示也类似。

（3）数据段（DATA）：以“DATA:”开始，以“ENDDATA”结束。为集合的属性（数组）输入必要的常数数据，格式为“attribute（属性）＝value_list（常数列表）；”。

常数列表（value_list）中数据之间可以用逗号“，”或空格分开（回车等价于一个空格），如上面对DEM的赋值也可以写成“DEM＝40 60 75 25；”。

在LINGO模型中如果需要在运行时为参数赋值，可以在数据段使用输入语句。但这仅用于为单个变量赋值，输入语句格式为“变量名＝？；”。例如，上例中如果需要在求解模型时才给出初始库存量（记为A），则可以在模型中数据段写语句“A＝？”。

在求解时LINGO系统给出提示界面等待用户输入变量A的数值，当然此时的约束语句 $INV(1)=10+RP(1)+OP(1)-DEM(1)$也应该改写成$INV(1)=A+RP(1)+OP(1)-DEM(1)$；。

这样模型可以计算任意初始库存量（而不仅仅只能计算初始库存量为10）的情况了。

（4）初始段（INIT）：以“INIT:”开始，以“ENDINIT”结束，为集合的属性（数组）定义初值（因为求解算法一般是迭代算法，所以如果用户能给出一个比较好的迭代初值，则对提高算法的计算效果有益）。

如果有一个接近最优解的初值，对LINGO求解模型也有帮助，定义初值的格式为“attribute（属性）＝value_list（常数列表）；”，这与数据段中的用法类似，帆船生产计划案例还没有此用法。

（5）计算段（CALC）：以“CALC：”开始，以“ENDCALC”结

束，对原始数据进行计算处理。

在实际问题中输入的数据通常是原始数据，不一定能在模型中直接使用，可以在这个段对这些原始数据进行一定的预处理，得到模型中真正需要的数据。如上例，如果希望得到全年的总需求和季度平均需求，可以增加这个段：

```
CALC：
    T_DEM = @SUM（TIMES： DEM）；           !总需求；
    A_DEM = T_DEM / @size（TIMES）；       !平均需求；
ENDCALC
```

在计算段中也可以使用集合函数（其中函数@size(quarters)表示集合quarters的元素个数，这里是4）。这时变量*T_DEM* 的值是总需求，*A_DEM* 的值是平均需求（如果需要，这两个变量可以在程序的其他地方作为常数使用）。

上面的两个语句不能交换顺序，因为计算*A_DEM* 必须要用到*T_DEM* 的值；此外在计算段中只能直接使用赋值语句，而不能包含需要经过解方程或求解优化问题后才能决定的变量。

案例：造完帆船造飞机——**Crystal Ball**求解

Bollinger公司主要为一家飞机引擎制造商生产两种不同的电子组件，即322A和802B，飞机引擎制造商在接下来的3个月里都会通知该公司的销售人员每个星期的需求量。每个月对组件需求量的变化可能很大，这要视飞机引擎制造商所生产的引擎而定。表2-9所示的是刚刚接到的未来3个月的订单。

表**2-9**　未来**3**个月的订单

组　　件	**4**月	**5**月	**6**月
322A	1 000	3 000	5 000
802B	1 000	500	3 000

接到订单后需求报告将会送到生产控制部，该部必须制订未来3个月的组件生产计划，在制订生产计划时需要考虑总生产成本、库存持有成本和改变生产负荷导致的费用。

根据以往的数据，生产一个322A组件的成本为$20；生产一个802B组件的成本为$10。每月的基本库存持有成本占生产成本的1.5%，即322A组件的库存成本为$0.3 / 组件；802B组件的库存成本为$0.15/组件。成本与生产负荷的水平有关系，生产负荷增加一个单位时新增的成本为$0.5；生产负荷减少一个单位时成本减少$0.2。假定3月份322A的生产量为1 500，802B的生产量为1 000。

322A组件的初期库存为500，6月最小安全库存为400；802B的初期库存为200，6月最小安全库存200，如表2-10所示。

表**2-10**　生产**322A**和**802B**的需消耗的机器、人工和占用的库容

产　　品	机　　器	人　　工	库　　存
322A	0.1	0.05	2
802B	0.08	0.07	2

如我们前面所讲，员工的工作时长、机器的最大生产能力和库容等因素均是限制生产的约束条件，如表2-11所示。

表**2-11**　机器生产能力、人工及最高库存能力限制

月　　份	机器能力	人　　工	库存最高

四月	400	300	10000
五月	500	300	10000
六月	600	300	10000

生产计划部的人员需要在以上条件下制订出满足客户的需求的最优生产计划方案。

（1）将信息抽象成数学模型

在满足客户需求的基础上生产计划的安排会直接影响总成本的变化，成本最优是公司的目 标，因此设置总成本为预测（目标）变量。

我们要做的决策是如何来安排生产计划，因此决策变量是每月322A和802B两种产品在未来3个月的生产计划，共6个决策变量。

该模型中共有4个约束条件，即6月末最小安全库存限制、每月总库容限制、机器和人工的能力限制。由于有些约束条件涉及每个月，因此每项约束条件展开可能有多项，如机器能力限制3个月就应是3个约束条件。

由于该例中未来几个月的订单数量已知，因此本模型中暂时不需要假设变量。

（2）建立Excel模型

总成本由3方面组成：一是生产成本，即生产所产生的成本；二是库存成本，即存放在仓库产生的成本；三是生产负荷变化成本，即每个月与上月的负荷变量导致的成本的变化。因此总成本与生产计划息息相关，逻辑关系如下。

·总成本＝生产成本＋库存成本＋变化成本。

·生产成本＝每月322A组件生产件数×单位成本＋每月802B组件

生产件数×单位成本。

·库存成本＝每月332A库存件数×库存单位成本＋每月802B库存件数×库存单位成本（每月库存＝上月末库存＋本月生产量-本月发货量）。

·变化成本＝（本月生产总量-上月生产总量）×变化单位成本。这里要注意如果是减少，则总量差为负值，再乘以变化单位成本。即总量减少该项成本对总成本有贡献，因此整理需要在Excel中通过IF函数进行逻辑判断。

在Excel中整理的模型如图2-81所示。

交货计划

组件	四月	五月	六月
322A	1000	3000	5000
802B	1000	500	3000

生产计划

组件	三月	四月	变化量	五月	变化量	六月	变化量
322A	1500	500		3200		5200	
802B	1000	2500	500	2000	2200	500	500

	期初库存	安全库存	四月库存	五月库存	六月库存
322A	500	400	0	200	400
802B	200	200	1700	3200	700

	四月	五月	六月
变化成本	250	1100	250
生产成本	228000		
库存成本	1020		
总成本	230620		

图2-81　Excel数学模型

打开“飞机部件生产计划.xls”工作表，设置说明如下。

·单元格H2:K4区域设置为未来3个月的交货计划，即市场的需求量。

·单元格H7:K9区域为未来3个月的生产计划，即决策变量区域。由于要考虑生产负荷变量对成本的影响，因此需要根据交货计划和生产计划计算出每月的变化量。在K9单元格中输入公式“＝J8＋J9-I8-I9”，其他月份依此类推。各决策变量的初值设置见Excel模型。

·单元格H11:M13区域用于计算库存变化量。

·单元格H15:K19区域用于计算生产总成本，即预测变量区域；除此之外，我们还缺少约束条件。为此在单元格B3:F19区域中增加模型的限制条件，更新后的Excel模型如图2-82所示。

成本变化情况

	生产成本	库存成本	负荷变化成本	
	单位成本	单位成本		
322A	20	0.30	增加	0.5
802B	10	0.15	减少	0.2

	机器	人工	库存
322A	0.1	0.05	2
802B	0.08	0.07	2
四月	400	300	10000
五月	500	300	10000
六月	600	300	10000

	机器	人工	库存
四月	250	200	6000
五月	480	300	10400
六月	560	295	11400

交货计划

组件	四月	五月	六月
322A	1000	3000	5000
802B	1000	500	3000

生产计划

组件	三月	四月	变化量	五月	变化量	六月	变化量
322A	1500	500		3200		5200	
802B	1000	2500	500	2000	2200	500	500

	期初库存	安全库存	四月库存	五月库存	六月库存
322A	500	400	0	200	400
802B	200	200	1700	3200	700

	四月	五月	六月
变化成本	250	1100	250
生产成本	228000		
库存成本	1020		
总成本	230620		

图2-82　更新后的Excel模型

该模型中的逻辑关系为：根据交货计划→确定生产计划→确定每月库存→确定总成本。

（3）设置模型中的各参数并计算

选择单元格J8和J9，设置决策变量参数，如图2-83所示。

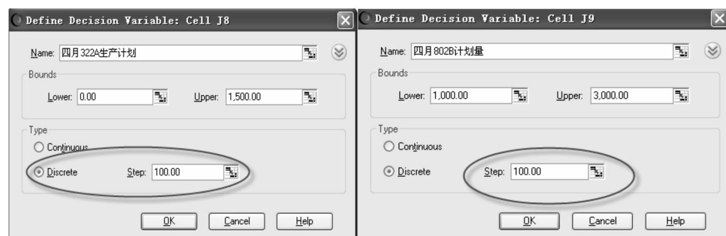


图2-83　设置决策变量

322A的生产计划设置为*Lower* ＝0，*Upper* ＝1500；802B的生产计划设置为*Lower* ＝1000，*Upper* ＝3000。

注意设置决策变量时，*Lower* 和*Upper* 分别指该变量的最小值和最大值，即在运算过程中决策变量的取值范围不会超过该区间。该区间的设置主要根据该参数的实际情况或经验而定，其范围值也可理解为决策变量的一种约束条件。该范围设置适当不仅可以减少模型的运行次数和时间，还可以提高模型运算的准确性。预测变量计算结果与实际情况有偏差时，可能是该决策变量的计算范围设置不恰当。

在【Type】选项组中选择【Continuous】（连续类型）单选按钮，按照蒙特卡洛抽样方式对决策变量进行抽样计算；选择【Discrete】单选按钮，会按照指定的步长计算，本例以100为一个步长计算。

选择Excel模型表中I19单元格设置为预测变量，如图2-84所示。

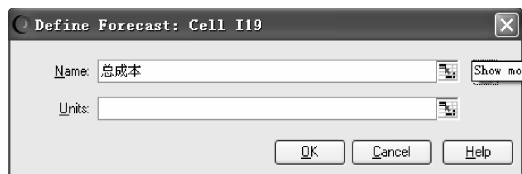


图2-84　设置目标函数

单击Crystal Ball菜单中【Tools】|【OptQuest】选项，显示【OptQuest】窗口。选择【Minimize】选项最小化预测变量“总成本”。如图2-85所示。

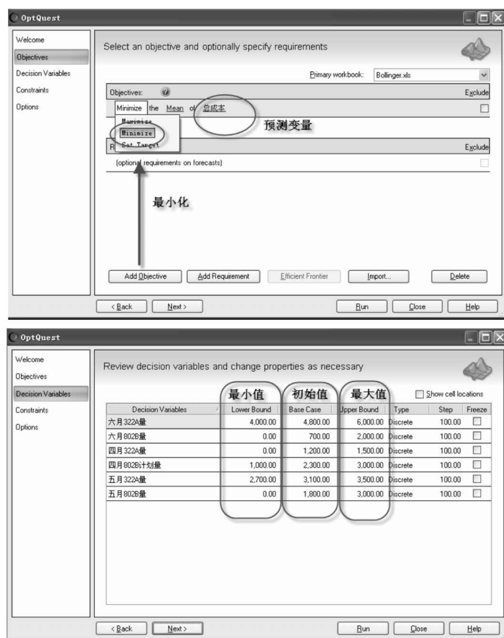


图2-85　设置优化计算

此时窗口中会显示决策变量的相关信息，包括最小值、最大值和初值。初值一定要在范围内，否则运算会有错误提示。

如果需要某个决策变量暂不参与运算，则选择其对应的【Freeze】复选框。

单击【Next】按钮，设置限制条件。共设置11个限制条件，如图2-86所示。

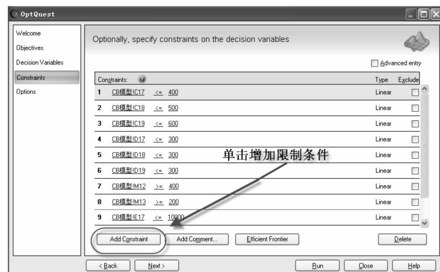
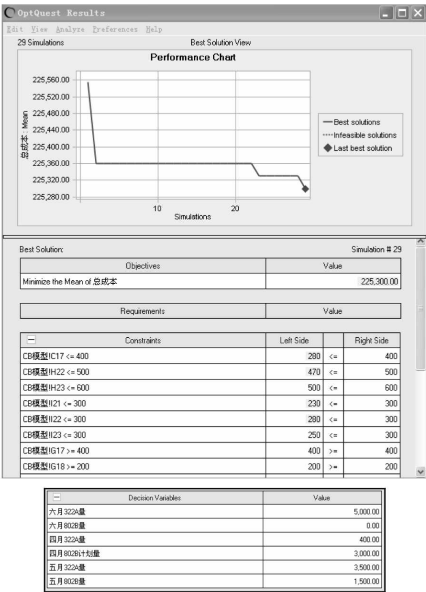


图2-86　设置限制条件

根据人工时、机器能力及库存对模型进行限制，如C17单元格表示4月份的机器能力，按照已知条件不能超过400，单击【Add Constraints】按钮，设置C17单元格＜＝400。其他约束条件依次类推。

单击【Next】按钮，单击【Run】按钮开始计算模型，计算结果如图2-87所示。



Objectives	Value
Minimize the Mean of 总成本	225,300.00

Requirements	Value

Constraints	Left Side		Right Side
CB模型!C17 <= 400	280	<=	400
CB模型!H22 <= 500	470	<=	500
CB模型!H23 <= 600	500	<=	600
CB模型!I21 <= 300	230	<=	300
CB模型!I22 <= 300	280	<=	300
CB模型!I23 <= 300	250	<=	300
CB模型!G17 >= 400	400	>=	400
CB模型!G18 >= 200	200	>=	200

Decision Variables	Value
六月322A量	5,000.00
六月802B量	0.00
四月322A量	400.00
四月802B计划量	3,000.00
五月322A量	3,500.00
五月802B量	1,500.00

图2-87　优化计算结果

可以看出在满足所有约束条件下，预测变量“总成本”的最小值接近225 300元。

Miss　Ju："对比这两个制造类案例，看起来使用LINGO需要有比较强的数学抽象思维。能够将有相同属性的变量进行集合。而Crystal　Ball好像更多的是需要逻辑思维，只要思路正确，在Excel中建立好模型就可以。"

Mr　Shu："不错，这两款软件各有优点，就看哪款软件适合你。"

Miss　Ju："可是我还是觉得LINGO在集合建模时比较难，可能Crystal Ball更适合我一些。"

Mr Shu："我个人认为Crystal Ball可能更适合初学者一些。"

Miss Ju："Crystal Ball还有没有什么好玩的一些功能？"

Mr Shu："我们前面介绍了它的预测和优化求解的部分功能，我们再看看它的一些其他应用"。

在前面的例子中有的是假设变量加预测变量求解风险概率值，有的是决策变量加预测变量求最优值，可不可以将假设变量加决策变量加预测变量组合在一个模型中？

案例：怎么进货才能赚得最多——书店运营（**Crystal Ball**求解）

在8月份书店需要确定订购的明年日历的数量，单本日历的进价是7.5美元且售价是10美元，2月份之后所有未售的日历将会以2.5美元的价格退还给出版商。为了保证利润的最大化，如何来确定明年的日历数量？店主将以往的日历的销售数据进行了分布拟合，其销售量在100本～300本之间（均匀分布）。

（1）理顺模型关系

·确定假设变量及预测变量

需求量＝d ~ Uniform［100,300］（呈均匀分布）。

订购量＝Q（随决定变化），则有关指标计算公式如下。

收入＝$10×Min（需求量，订货量）。

成本＝$7.50×$Q$。

退款＝$2.50×Max（定货量-需求量，0）。

利润＝收入-成本＋退款。

客户的需求量不可控，由此将需求量作为假设变量。而书店的利润为目标，故设置为预测变量。

·在Excel中设置假设变量及预测变量模型，如图2-88所示。

	A	B	C	D	E	F
1						
2	书店进书模型					
3	数据					
4	单位成本	$7.50				
5	单位售价	$10.00				
6	退货售价	$2.50				
7						
8	需求分布					
9	最小需求	100				
10	最大需求	300				
11						
12	模拟					
13	订货数量	需求量	收入	成本	退款	利润
14	100	200	$1,000.00	$750.00	$0.00	$250.00
15	150	200	$1,500.00	$1,125.00	$0.00	$375.00
16	200	200	$2,000.00	$1,500.00	$0.00	$500.00
17	250	200	$2,000.00	$1,875.00	$125.00	$250.00
18	300	200	$2,000.00	$2,250.00	$250.00	$0.00

图2-88　设置假设变量及预测变量模型

打开“书店运营风险.xls”工作表，设置如下。

在C14单元格中输入公式“＝$B$5*MIN(A14,B14)”，即销售日历

的收入，并将公式下拉至C18单元格。

在D14单元格中输入公式“＝A14*$B$4”，即购买日历的成本值，并将公式下拉至D18单元格。

在E14单元格中输入公式“＝MAX((A14-B14)，0)*$B$6”，即退款的数量，并将公式下拉至E18单元格。

在F14单元格中输入公式“＝C14-D14＋E14”，即所得利润，并将公式下拉至F18单元格。为了说明决策变量，暂时将订货数量设置为固定数。

选择B14单元格，单击【Crystal Ball】|【Define Assumption】|【Distribution Galley】选项，选择【Uniform】选项，显示【Define Assumption】窗口，如图2-89所示。

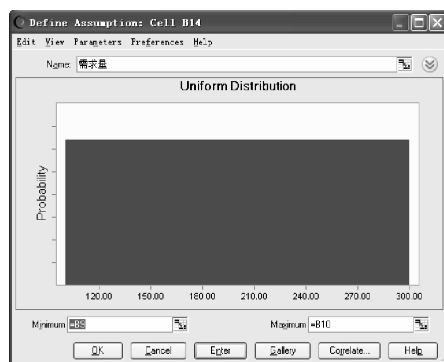


图2-89 【Define Assumption】窗口

变量名称设置为“需求量”，在【Minimum】下拉列表框中引用B9单元格，在【Maximum】下拉列表框中引用B10单元格内容。这样主要是为了方便在Excel中修改数值，B15:B18单元格区域依此进行设置。

选择F14单元格，单击【Define Forecast】选项。变量名设置

为“利润”，如图2-90所示。

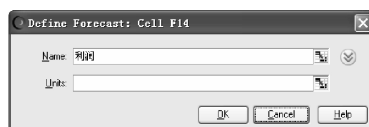


图2-90　设置预测变量

F15:F18单元格区域依此设置。

截至目前，我们在假设已知未来几月的订货数量的情况下建立了模型。

（2）预测书店利润

单击【Run】按钮计算，将运行结果以趋势图的形式呈现，如图2-91所示。

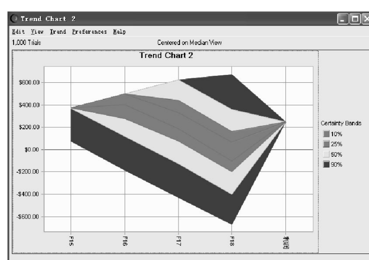


图2-91　运行结果

趋势图针对预测单元格给出了确定的频带，项目执行期间有10%的几率位于内带（浅蓝色）；25%位于第2带（红色）；50%位于第3带（绿色）；90%的几率位于外带（深蓝色）。

从图中可以看出F18单元格（最后一个月利润）出现负值的概率

最大，其中分别说明了未来几个月的利润有可能出现的概率分布情况。

（3）在Excel模型中加入决策变量

在此之前，我们在Excel模型中均假定未来几个月的订货数量已知。实际上，作为书店经营者往往最关心的就是："我该如何确定订货量？"

从Excel模型中可以看出订货量在200本时，该日历的利润可达到最大。这与日常的经营理念也相符，即能卖多少就进多少本，没有剩余最为精益，我们现在可以尝试通过加入决策变量看计算结果是否与上述结果相同。

打开"书店运营优化.xls"工作表，为了与上述方案对比，暂将需求量定为200。选择B14单元格，单击【Crystal Ball】选项，单击【Clear】按钮清除假设变量，如图2-92所示。

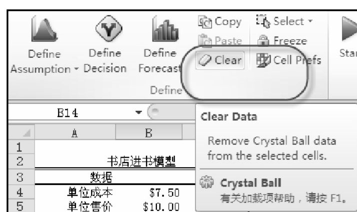


图2-92　清除假设变量

选择A14单元格，单击【Define Decision】选项设置决策变量，如图2-93所示。

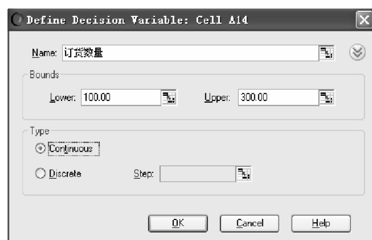


图2-93　设置决策变量

决策变量命名为“订货数量”，【Lower】设置为“100”，【Upper】设置为“300”，【Type】设置为“Continuous”，如图2-94所示。

	A	B	C	D	E	F
1						
2	书店进书模型					
3	数据					
4	单位成本	$7.50				
5	单位售价	$10.00				
6	退货售价	$2.50				
7						
8	需求分布					
9	最小需求	100				
10	最大需求	300				
11						
12	模拟					
13	订货数量	需求量	收入	成本	退款	利润
14	168	200	$1,680.00	$1,260.00	$0.00	$420.00

图2-94　设置变量

（4）设置优化计算

选择【Crystal　Ball】|【OptQuest】选项，在【OptQuest】窗口中单击【Add Objective】按钮。设置最大化目标，其他设置过程略。

单击【Run】按钮，计算结果如图2-95所示。

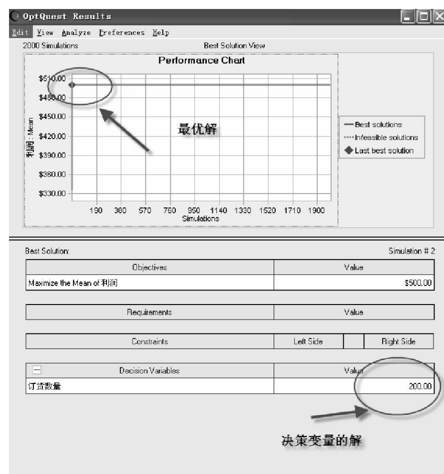


图2-95　优化计算结果

结算结果与Excel静态时计算结果一致，即在订货数量与需求量一致的情况下可以取得最大利润。但实际情况是需求量往往不是固定值，而是服从某种分布。即由于需求量往往是未知，这时我们应该如何来确定供货量？

（5）第2次添加假设变量

选择B14单元格，将其重新设置为假设变量（过程省略），修正后的模型如图2-96所示。

	A	B	C	D	E	F
1						
2	书店进书模型					
3	数据					
4	单位成本	$7.50				
5	单位售价	$10.00				
6	退货售价	$2.50				
7						
8	需求分布					
9	最小需求	100				
10	最大需求	300				
11						
12	模拟					
13	订货数量	需求量	收入	成本	退款	利润
14	145	200	$1,450.00	$1,087.50	$0.00	$362.50

图2-96　修正后的模型

（6）重新进行优化计算

以上模型有假设变量、决策变量和预测变量，与上次拟合不同的是需求量不再是定值，而是概率分布，计算结果如图2-97所示。

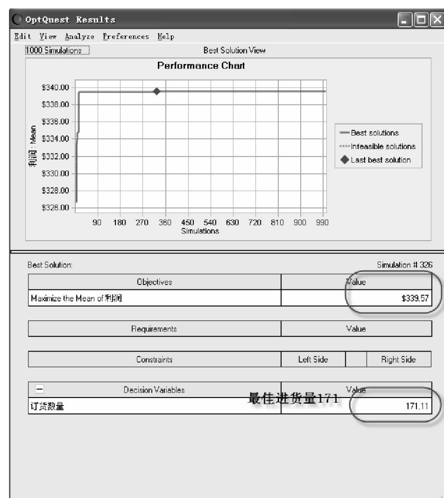


图2-97　修正后的优化计算结果

计算结果说明当订货数量在171（171.11取整数）本时，利润达到最大值$340。它低于前面计算的最高利润$500，主要是因为需求量有波动所致。这也是最符合实际情况的最优值，因为存在不确定性，所以实际情况下的最优值一般比理想情况下的最优值有所偏离。

案例：龙卷风来了，赶紧搬家（Crystal Ball求解）

Miss Ju：“这是什么意思？”

Mr　Shu：“这个案例用来展现Crystal　Ball在项目管理中的典型应用。”

Miss Ju："项目管理？Crystal Ball不是主要应用在风险分析和优化计算吗？在项目管理上也有应用吗？"

Mr Shu："工具的应用能否发挥到极致往往与使用者的思维有关，在项目管理中我们经常需要使用甘特图。然后使用图表寻找关键路径，利用Crystal Ball同样可以完成"。

Miss Ju："关键路径？能不能介绍一下项目管理的基本常识？"

Mr Shu："好的，在使用工具前首先介绍一下项目管理的基本概念"。

项目是为了特别需要并根据特定的目标在有限资源的约束下由一个专门组建的机构在事先设置的时间内实施一些相关联活动的总称，一项建筑工程的设计、一艘船舶的建造、一套设备的检修和一项研究计划的实施等都是项目。

关键路线法和计算评审技术是项目管理中最重要和应用最广泛的两项技术，前者可以在网络模型上直观地分析工程项目所需时间，找到缩短工程日期的关键所在。但其不足之处是各活动的持续时间是确定的，而实际上任何项目的大多数活动的持续时间通常都是不确定的。在项目尚未开始时活动的持续时间只能是一个估计值，在项目执行过程中持续时间将受各种因素的影响而发生变化；后者在关键路线的基础上将随机性引入活动的持续时间，项目的结束时间也具有随机性。再来评估项目结束时间是否在一定范围的概率，因此使技术较关键线路法更能准确地模拟项目过程。

根据项目管理的特点，有经验的项目管理人员往往能够估计一个活动的3个参数，即活动持续的最小值、最可能值和最大值。因此通常将活动时间假设服从三角分布，下面以某公司的搬迁计划为例说

明：

某公司计划将信用卡的经营地点从A地搬到B地，搬迁项目涉及公司多个不同部门。人事部要决定哪些员工从原地点调离？需要雇用多少新的员工？谁来负责新员工的培训？系统小组 和财务部必须制定新的操作程序并确定财政安排；设计师要进行室内设计并监管需要的结构改善工程，每个场所都是有足够开放空间的既存建筑物，必须拥有写字间、计算机设施和设备等。

问题的复杂性在于各项活动相互关联，即项目的一些部分必须在另一部分结束之后开始。例如，公司不能在室内设计之前就进入办公室内部，并且在人事需求没有确定之前不可能雇用新的员工。办公室搬迁项目顺序如图2-98所示。

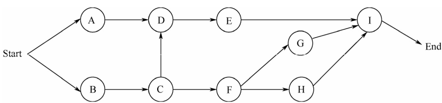


图2-98　办公室搬迁项目顺序

各项目的具体任务及可能的时间如表2-12所示。

表**2-12**　各项目的具体任务及可能的时间

项目	描述	前项工作	时间估计（天数）				
			最小值	最可能	最大值	开始	结束
A	办公室选址	—	19	21	22	0	21
B	建立组织机构和财政计划	—	20	25	30	0	25
C	确认个人需求	B	15	20	30	25	45
D	设计设施	A，　C	20	28	42	45	73
E	建设设施	D	40	48	66	73	121
F	前期人员迁移	C	10	12	13	45	57
G	招聘新员工	F	20	25	32	57	82
H	关键员工迁移	F	25	28	29	57	85
I	新人培训计划	E、G和H	10	15	24	121	136

现在该公司要想了解如何在最短的时间内完成所有任务？哪些任务是最关键的？操作步骤如下。

（1）将项目过程抽象成数学模型

打开“关键路径.xls”工作表，在Excel数学模型中设置的公式如表2-13所示。

表2-13　设置公式

最 小 值	最 可 能	最 大 值	开 始 时 间	结束时间
19	21	22	0	=F6+H6
20	25	30	0	=F7+H7
15	20	30	=H7+F7	=F8+H8
20	28	42	=MAX(H6+F6,H8+F8)	=F9+H9
40	48	66	=H9+F9	=F10+H10
10	12	13	=H8+F8	=F11+H11
20	25	32	=H11+F11	=F12+H12
25	28	29	=H11+F11	=F13+H13
10	15	24	=MAX(H10+F10,H12+F12,H13+F13)	=F14+H14

在单元格E6:G14区域设置每项活动持续时间的3个参数，其中A和B项活动不受其他活动的影响，可以同时进行；A/B项开始时间时间为0；C项活动必须在B项活动结束后才可以进行，因此C项活动的开始时间为B项活动的结束时间，即“F7＋H7”；D项活动必须在A和C项活动都结束后才开始，其开始时间为A/C项活动最晚结束的时间，即“＝MAX(H6＋F6,H8＋F8)”；其他活动时间依此类推。

（2）设置假设变量与预测变量

单元格F6:F14区域为假设变量区域，为各项持续时间的分布。预测变量为项目结束时间，各项目的时间分布均选择为三角形分布。以F6单元格设置假设变量为例，如图2-99所示。

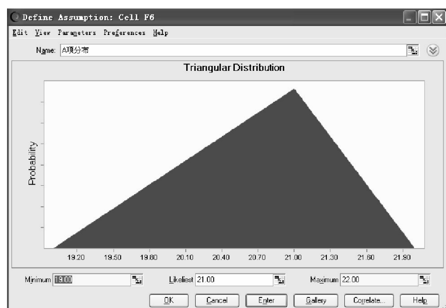


图2-99　设置假设变量

在Crystal Ball分布库中选择三角形分布，操作过程略。

在【Name】下拉列表框中输入“A项分布”，在【Minimum】下拉列表框中输入“19”，在【Maximum】下拉列表框中输入“22”，在【Likeliest】下拉列表框中输入“21”。分别表示A项活动需要消耗的时间最可能为21天，最大可能为22天，最小可能为19天。

其他项目的时间分布依次进行假设变量设置。

选择I14单元格，设置预测变量，如图2-100所示。

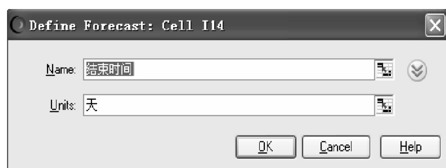


图2-100　设置目标函数

在【Name】下拉列表框中输入“结束时间”，在【Units】下拉列表框中输入“天”。

（3）模型运算

设置假设变量及预测变量后单击【start】按钮，模型运算结果如

图2-101所示。

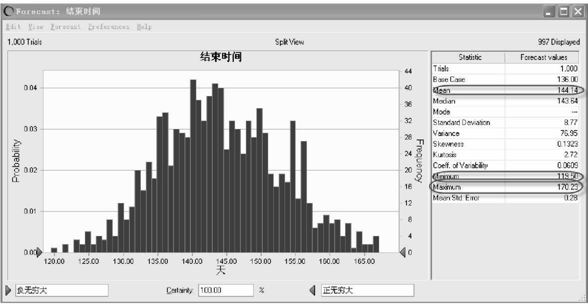


图2-101　模型运行结果

从计算结果可以看出本项目最长时间为170天，最短为119天，平均144天。少于140天内完成该项目的概率为33.39%，大于140天内完成该项目的概率为66.61%，如图2-102所示。

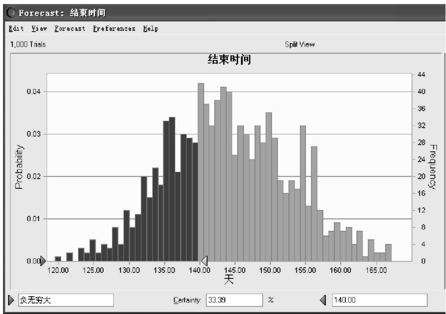


图2-102　140天以内完成该项目概率

（4）敏感性分析

在【Crystal　Ball】菜单中选择【View　Charts】|【Sensitivity Charts】选项，显示【Sensitivity：结束时间】窗口。其中的敏感度图给出了随机变量（活动次数）对输出结果（项目完成时间）的影响程度，如图2-103所示。

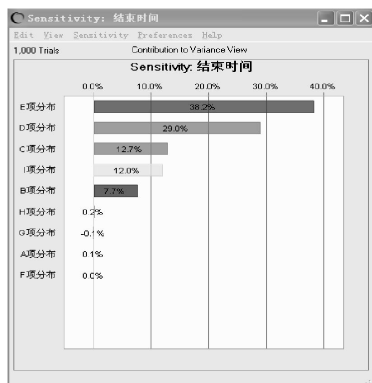


图2-103　敏感度图

活动E的变化对整个项目的持续时间影响最大，然后依次是活动D、C、I和B。活动G的 变化对项目几乎没有影响，因此可以看出缩短项目时间的关键活动是E和D。

除了敏感性分析可以看出各项活动对项目的影响外，龙卷风图分析还可以定量分析每项活动对项目时间的影响。

（5）龙卷风图分析

单击【More tools】|【Tornado Chart】选项，显示【Specify target】对话框，如图2-104所示。

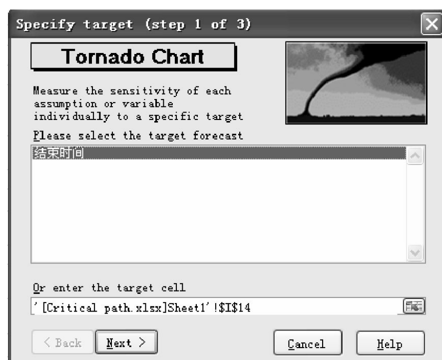


图2-104　【Specify target】对话框

单击【Next】按钮，添加假设变量，如图2-105所示。

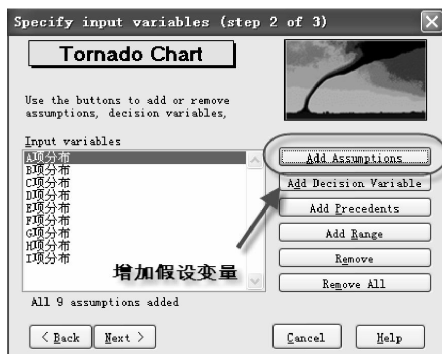


图2-105　添加假设变量

将A～I项的任务添加至假设变量中，单击【Next】按钮后设置范围，如图2-106所示。

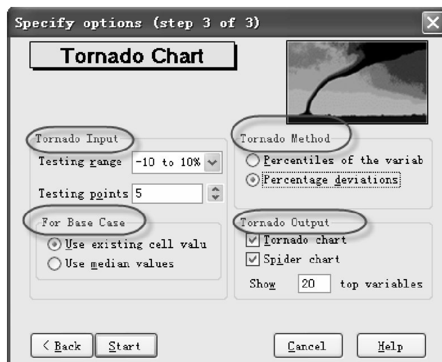


图2-106　设置范围

不同的设置导致的计算结果大同小异，简要说明如下。

·【Tornado　Input】选项组：主要用于设置变量的范围和测试的次数，本次计算设置为10%且计算5次。

·【Tornado Method】选项组：用于选择分析方法，选择【Percentage deviations】单选按钮，为百分比偏差方法；选择【Percentiles of the variab】单选按钮，为变量的百分位数方法，本次计算选择前者。

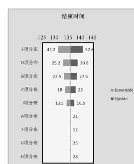


图2-107　龙卷风图的计算结果

·【For Base Case】选项组：用于设置计算的基点数据，选择【USe existing cell value】单选按钮，为单元格中默认的数字；选择【Use median values】单选按钮，用于中位数，本次计算选择前者。

·【Tornado Output】选项组：用于设置需要的图形和变量，选择【Tornado chart】单选按钮，为龙卷风图；选择【Spider chart】单选按钮，为蜘蛛图；同时指定前20位最显著的影响变量。

单击【Start】按钮，计算结果会自动在两个新建的Excel表中呈现，龙卷风图的计算结果如图2-106所示。

龙卷风图的分析结果如表2-14所示。

表**2-14**　龙卷风的分析结果

	结束时间			Input		
Variable	Downside	Upside	Range	Downside	Upside	Base Case
E 项分布	131.2	140.8	9.6	43.2	52.8	48
D 项分布	133.2	138.8	5.6	25.2	30.8	28
B 项分布	133.5	138.5	5	22.5	27.5	25
C 项分布	134	138	4	18	22	20
I 项分布	134.5	137.5	3	13.5	16.5	15
A 项分布	136	136	0	18.9	23.1	21
F 项分布	136	136	0	10.8	13.2	12
G 项分布	136	136	0	22.5	27.5	25
H 项分布	136	136	0	25.2	30.8	28

在基础数据上下浮动10%的计算结果，如E项分布的基点数据为48，上浮10%为52.8。如此时将Excel模型中的E项分布的初值修改为52.8，则根据模型计算预测变量“结束时间”为140.8，即为UpSide值。计算表格已按照各变量的影响大小排序，运算结果与敏感性分析结果一致。

除了可以运行龙卷风分析外，选择蜘蛛图的计算结果如图2-108所示。

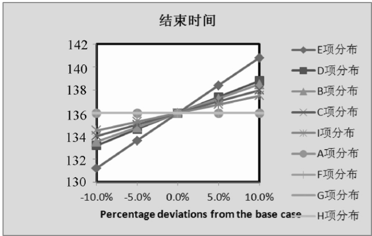


图2-108　蜘蛛图的计算结果

蜘蛛图的分析结果如表2-15所示。

表**2-15**　蜘蛛图的分析结果

	结束时间				
Variable	−10.0%	−5.0%	0.0%	5.0%	10.0%
E 项分布	131.2	133.6	136	138.4	140.8

续表

	结束时间				
D 项分布	133.2	134.6	136	137.4	138.8
B 项分布	133.5	134.75	136	137.25	138.5
C 项分布	134	135	136	137	138
I 项分布	134.5	135.25	136	136.75	137.5
A 项分布	136	136	136	136	136
F 项分布	136	136	136	136	136
G 项分布	136	136	136	136	136
H 项分布	136	136	136	136	136

通过蜘蛛图中各线的斜率也能看出各假设变量对预测变量的影响程度，与龙卷风图的分析结果一致；同时将变量上下浮动10%均分为5次计算。

除了按照单个变量的百分比偏差计算外，还可以选择Percentiles of the variab（变量的百分位数）计算方法，即按照分布的区间计算。

计算设置如图2-109所示。

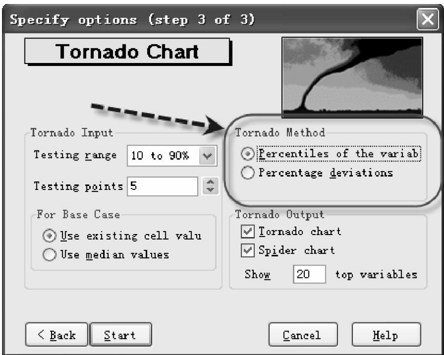


图2-109　按变量百分数计算设置

此时选择按照分布的区间计算，测试的范围设置为10%～90%，按照三角形分布的范围计算。以E项活动为例，通过计算机拟合最小值为40、最可能值为48和最大值为66的三角形分布在左右尾概率为

0.1的情况下，如图2-110所示。

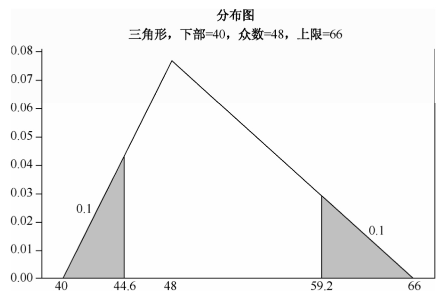


图2-110　三角形概率分布

在计算过程中E项活动的范围为44.6 ~ 59.2，单击【Start】按钮，计算结果如图2-111所示。

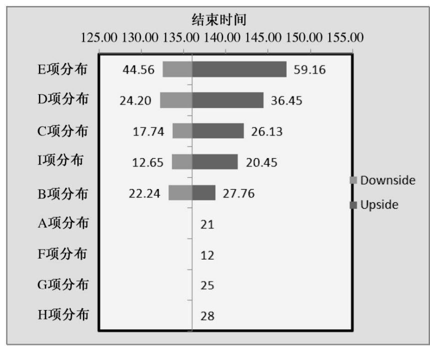


图2-111　计算结果

龙卷风的分析结果如表2-16所示。

表2-16　龙卷风的分析结果

	结束时间			Input		
Variable	Downside	Upside	Range	Downside	Upside	Base Case
E 项分布	132.56	147.16	14.60	44.56	59.16	48
D 项分布	132.20	144.45	12.25	24.20	36.45	28

续表

	结束时间			Input		
Variable	Downside	Upside	Variable	Downside	Upside	Variable
C 项分布	133.74	142.13	8.39	17.74	26.13	20
I 项分布	133.65	141.45	7.80	12.65	20.45	15
B 项分布	133.24	138.76	5.53	22.24	27.76	25
A 项分布	136	136	0	19.77	21.45	21
F 项分布	136	136	0	10.77	12.45	12
G 项分布	136	136	0	22.45	29.10	25
H 项分布	136	136	0	26.10	28.37	28

计算结果“Input”变量中E项的计算范围与概率分布模拟的一致（44.6～59.2），从以上结果可以看出在利用分布进行计算时预测变量的变化范围会更大。如果在计算设置时，将概率范围设置为20%～80%，则预测变量计算的范围也会相应变窄。

前述步骤中的【Testing Range】范围由“10to90%”修正为“20to80%”，则蜘蛛图计算结果如图2-112所示。

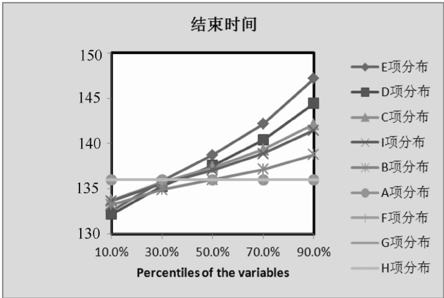


图2-112　修正后蜘蛛图的计算结果

修正后的蜘蛛图的分析结果如表2-17所示。

表**2-17**　修正后的蜘蛛图的分析结果

			结束时间		
Variable	10.0%	30.0%	50.0%	70.0%	90.0%
E 项分布	132.5607017	135.8993671	138.7029415	142.1509494	147.1589474

续表

			结束时间		
Variable	10.0%	30.0%	50.0%	70.0%	90.0%
D 项分布	132.1952354	135.2663608	137.5903264	140.3875081	144.4502252
C 项分布	133.7386128	135.7434165	137.339746	139.2917961	142.1270167
I 项分布	133.6457513	135.5825757	137.0627461	138.8518295	141.4503521
B 项分布	133.236068	134.8729833	136	137.1270167	138.763932
A 项分布	136	136	136	136	136
F 项分布	136	136	136	136	136
G 项分布	136	136	136	136	136
H 项分布	136	136	136	136	136

当然，我们还可以尝试利用中位数进行计算，这里就不再一一讲解。使用中位数计算时，*Input* 变量的计算范围不变，上下限也不变。预测变量的波动范围不变，上下限数值有变化。

从以上计算来看，无论使用何种计算方法得出的定性的结论都是一致的，使用龙卷风图可以更清晰地看出E、D、C和I等关键因素对目标变量的影响范围。

Miss Ju："概率分布还可以应用在项目管理上，真的是很神奇。"

Mr Shu："所以说困难有没有，关键看思路，思路对了，困难自然迎刃而解。我们对这一章的内容做一个简短的回顾吧。"

从数鸡兔开始，我们了解了Excel的规划求解功能，然后分析了几个典型的规划求解的案例。之后又了解了Excel的其他数据分析功能，如模拟计算、方案管理器及单变量求解等。

在知道Excel的规划求解功能后，我们从农夫种田的案例开始使用图解法了解线性规划的基本原理。非线性或整数线性规划使用图解法比较困难，需要使用Crystal Ball来完成，还可以使用LINGO软件来计算，这两款软件各有优点，LINGO需要较强的数学思维；Crystal Ball可以结合Excel的强大功能。

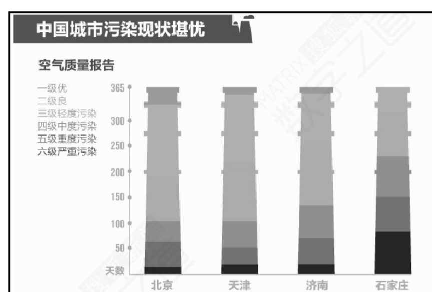
在上节中列举了多个Crystal Ball的应用案例，从中可以了解其在风险与优化的综合应用，以及在项目管理上应用。

第3章
图个明白，精彩展现
JMP精彩图表

3.1　图个明白——常用图形

3.2　图个明白——树图

3.3　图个明白——SPC图



Mr Shu："知道制作PPT的基本原则是什么吗？"

Miss Ju："你指的是文字、数据和图表的关系吗？"

Mr Shu："是的。"

Miss Ju："一般来讲，优先顺序应该是图表、数据和文字，能用图表说明的就不用数据，能用数据说明的就不用文字，对吗？"

Mr Shu："是的，图表是非常好的沟通工具，也是数据分析结果展现的重要工具。你看上面的图形能得出什么结论？"

Miss Ju："北京相对于其他3个城市的空气能稍好，石家庄最差。就算是相对最好的北京的空气质量在3级污染以下的时间也长达200天，空气质量堪忧啊。"

Mr Shu："开始忧国忧民了？是的，图表的确非常清楚。Excel、Minitab和Crystal Ball都有很多图表可以使用，如Minitab与JMP中的主要图形如图3-1所示。

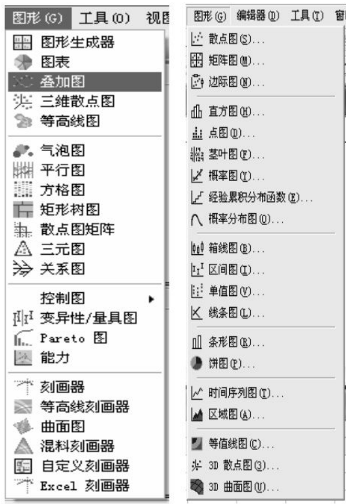


图3-1　JMP及Minitab的主要图形

相比Minitab软件，JMP的图形更加丰富多彩。下面我们重点介绍

几个常用的图形，主要通过Minitab和JMP软件来操作。

由SAS（全球著名的统计学软件公司）推出的一种交互式可视化统计发现软件系列包括JMP、JMP Pro、JMP Clinical、JMP Genomics和SAS Simulation Studio for JMP等强大的产品线，主要用于实现统计分析。JMP的算法源于SAS，特别强调以统计方法的实际应用为导向。其交互性和可视化能力强，使用方便。尤其适合非统计专业背景的数据分析人员使用，在同类软件中有较大的优势。

与Minitab一样，JMP的应用领域主要在数据统计分析上包括业务可视化、探索性数据分析、六西格玛及持续改善（可视化六西格玛、质量管理和流程优化）、试验设计、生存及可靠性、统计分析与建模、交互式数据挖掘，以及分析程序开发等。JMP是六西格玛软件的鼻祖，当年摩托罗拉开始推六西格玛时就用JMP软件。目前非常多的全球顶尖企业采用JMP作为六西格玛软件，包括陶氏化学、惠而浦、铁姆肯、招商银行、美国银行和中国石化等。

在医药领域，以严格和严谨著称的美国食品与药物管理局（FDA）对于药企申报的新药报告中的统计分析部分只接受使用SAS和JMP分析得出的统计结果，其40%以上的药物评审员都是JMP用户。

3.1　图个明白——常用图形

1．常用图形——Pareto及饼图

操作步骤如下。

（1）模拟随机数据

随机模拟一组银行客户数据，在Minitab菜单中选择【计算】|

【产生模块化数据】|【文本值】选项，如图3-2所示。

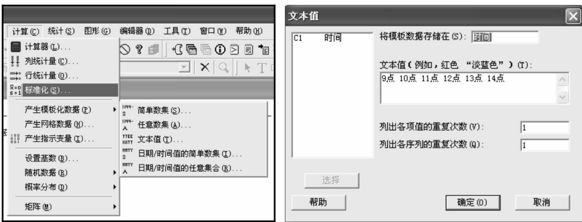


图3-2 模拟随机数据操作

在【将模板数据存储在】文本框中输入"时间"，在【文本值】下拉列表框中输入“9点10点11点12点13点14点”，各文本值之间键入空格，值和序列的重复次数均设置为1。单击【确定】按钮，在Minitab数据表格中出现随机产生的模拟时间数据，如图3-3所示。

工作表 1 ***

↓	C1-T	C2	C3
	时间		
1	9点		
2	10点		
3	11点		
4	12点		
5	13点		
6	14点		

图3-3 Minitab模拟数据

在Minitab菜单中选择【计算】|【产生模块化数据】|【任意数集】选项，显示【任意数集】对话框，如图3-4所示。

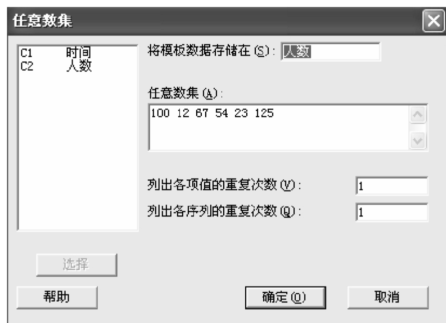


图3-4 【任意数集】对话框

在【将模板数据存储在】文本框中输入“人数”，在【任意数集】下拉列表框中输入“100 12 67 54 23 125”等，各数字之间键入空格，值和序列的重复次数均设置为1。

单击【确定】按钮，结果如图3-5所示。

工作表 1 ***

↓	C1-T	C2	C3
	时间	人数	
1	9点	100	
2	10点	12	
3	11点	67	
4	12点	54	
5	13点	23	
6	14点	125	

图3-5 Minitab模拟结果

（2）制作饼图

在Minitab菜单中选择【图形】|【饼图】选项，显示【饼图】对话框，如图3-6所示。



图3-6 【饼图】对话框

选择【用整理好的表格画图】单选按钮，在【类别变量】列表框中选择【时间】列，在【汇总变量】列表框中选择【人数】选项。单击【标签】按钮，显示【饼图-标签】对话框，如图3-7所示。在【标签扇形区】选项组中选择【类别名称】和【频率】复选框，单击【确定】按钮，显示的饼图，如图3-8所示。



图3-7 【饼图-标签】对话框

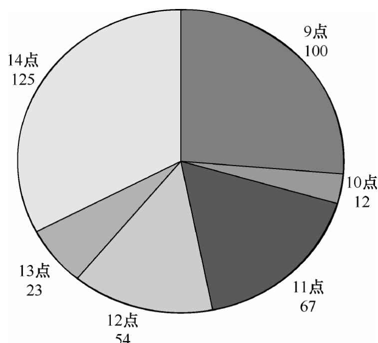


图3-8　Minitab饼图

（3）制作Pareto图

Pareto图大部分用来表示造成质量缺陷的频次图，即显示一个与质量相关制程或操作中问题的相对频数或严重性。该图是条形图，显示以降序排列的问题的分类，这里我们仍然以银行客户数据为例说明。

在Minitab软件中选择【统计】|【质量工具】|【Pareto图】选项，显示【Pareto图】对话框。选择【已整理成表格的缺陷数据】单选按钮，标签位于“时间”列，频率位于“人数”列，如图3-9所示。

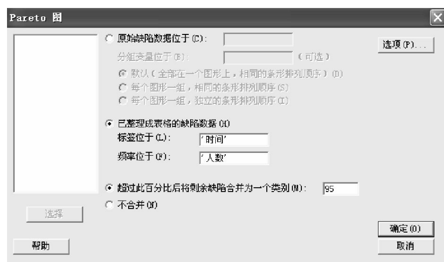


图3-9　设置Pareto图

单击【确定】按钮，结果如图3-10所示。

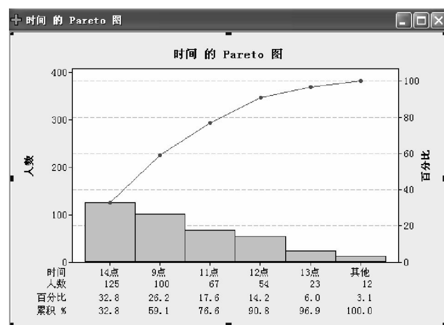


图3-10　Pareto图

Pareto图会自动根据数字的大小对柱状排序，红色线条是累计百分比线，帮助判断每个类别所加入的贡献。

JMP软件的操作步骤如下。

（1）复制数据至JMP工作表

将Minitab中的数据复制至JMP数据表中，如图3-11所示。

	时间	人数
1	9点	100
2	10点	12
3	11点	67
4	12点	54
5	13点	23
6	14点	125

图3-11　JMP数据表

（2）制作Pareto图

在JMP菜单中选择【图形】|【Pareto图】选项，显示【Pareto图】窗口，如图3-12所示。



图3-12　【Pareto图】窗口

在【Y，原因】下拉列表框中选择【时间】列，在【频数】列表框中选择【人数】列。单击【确定】按钮，完成的Pareto图如图3-13所示。

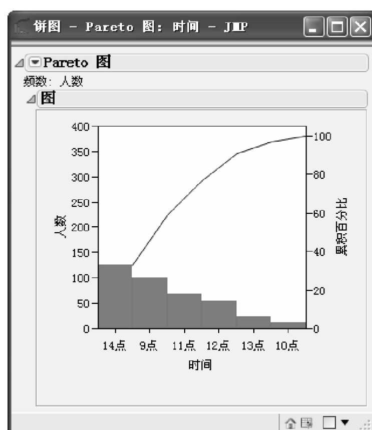


图3-13　完成的Pareto图

（3）制作饼图

单击已完成的Pareto图左上角的三角形，单击【饼图】选项即可将Pareto图转换成饼图，如图3-14所示。

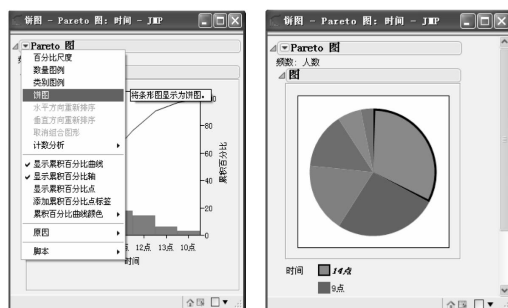


图3-14　Pareto图转换为饼图

在JMP中也可以要求生成对比Pareto图，即组合同一*Y*　变量的两个或多个Pareto图的图形显示。Pareto图可为指定为*X*　角色的每个值或者为来自两个*X*　变量的水平组合生成单独的图形显示，指定*X*　角色的列称为“分类变量”。

Pareto图命令可以将单个*Y*　（制程）变量制作为不带*X*　分类变量、带单个*X*　变量或带两个*X*　变量的图形。Pareto工具不区分数值变量和字符变量，不同于建模类型。所有值都被视为离散的，而条柱表示计数或百分比。Pareto图形生成的特点如下。

·不带*X*　分类变量的*Y*　变量将生成单个图表，其中*Y*　变量的每个值都有一个条柱。

·带一个*X*　分类变量的*Y*　变量将生成一行Pareto图。*X*　变量的每个水平都有一个图形，其中每个*Y*　水平都有条柱。

·带两个*X*　变量的*Y*　变量将生成多行和多列Pareto图，列出的第1个*X*　变量的每个水平都有一行；列出的第2个*X*　变量的每个水平都有一列。如前所述，对于第1个*X*　变量的每个值，这些行都有一个Pareto图。

（4）制作对比Pareto图

修改银行客户人数数据表，如图3-15所示。

	时间	人数	区间
1	9点	100	第一天
2	10点	12	第一天
3	11点	67	第一天
4	12点	54	第一天
5	13点	23	第一天
6	14点	125	第一天
7	15点	110	第一天
8	16点	76	第一天
9	17点	12	第一天
10	18点	34	第一天
11	9点	56	第二天
12	10点	67	第二天
13	11点	12	第二天
14	12点	34	第二天
15	13点	56	第二天
16	14点	109	第二天
17	15点	110	第二天
18	16点	43	第二天
19	17点	76	第二天
20	18点	23	第二天

图3-15　修改后的数据表

将数据表中时间延长至第2天，并新增“区间”列标示在“第一天”，还是“第二天”。在JMP中选择【图形】|【Pareto图】选项，显示【Pareto图】窗口，如图3-16所示。



图3-16　【Pareto图】窗口

在【Y，原因】下拉列表框中选择【时间】选项，在【X，分组】下拉列表框中选择【区间】选项。单击【确定】按钮，结果如图3-17所示。

这两个图形的水平轴和垂直轴的尺度相同，第1个图形的条柱以*y*轴值的降序排列，并确定所有单元的顺序。

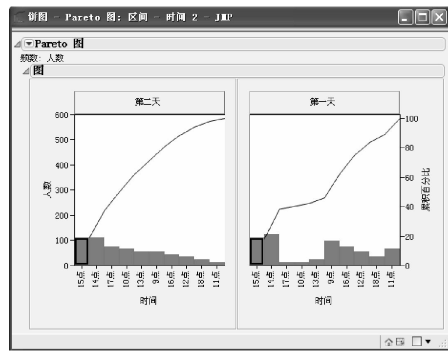


图3-17　增加*X* 分组后的Pareto图

2．常用图形——条形图、线形图及点图

测量一组11岁～16岁学生的相关信息，如身高和体重，共有233个学生参与检测。打开“学生.JMP”文件，相关信息如图3-18所示。

	年龄	性别	体重	身高
1	11	F	85	126
2	11	F	69	127
3	11	F	69	124
4	11	F	104	132
5	11	F	51	121
6	11	F	85	132
7	11	F	81	124
8	11	F	96	128
9	11	F	64	123
10	11	F	84	126
11	11	F	77	130
12	11	F	78	128

图3-18　相关信息

设置条形图的操作步骤如下。

（1）启动图表命令

在JMP中选择【图形】|【图表】选项，显示【图表】窗口，如图

3-19所示。



图3-19 【图表】窗口

在【选项】选项组中可选择图表的方向，即垂直或水平。还可以选择图表类型，即条形图、线形图、饼图、针状图和点图，在显示图表后仍可以随时更改这些选项。

在【类别，X，水平】列表框中最多可以选择使用两个*X* 变量，其水平将作为X轴上的类别。【图表】窗口可以为*X* 变量的每个水平或水平组合生成一个条形柱，如果没有指定*X* 变量，则数据表中的每一行都将在图表中显示为一个条形柱。

（2）设置条形图

在【类别，X，水平】列表框中选择【身高】选项作为*X* 轴，保留【统计量】列表框的默认设置，如图3-20所示。



图3-20　设置条形图

单击【确定】按钮，生成的条形图如图3-21所示。

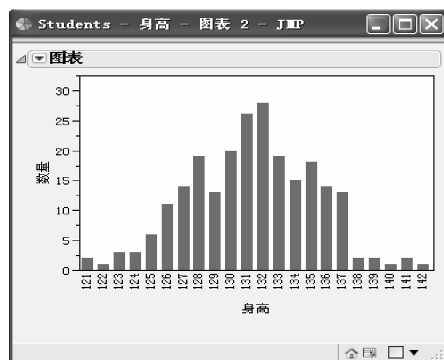


图3-21　生成的条形图

X 轴为“身高”，*Y* 轴为身高区间的统计量形成一个身高直方图。*Y* 轴默认统计量为数量，还有很多其他统计量可以选择，如图3-22所示。

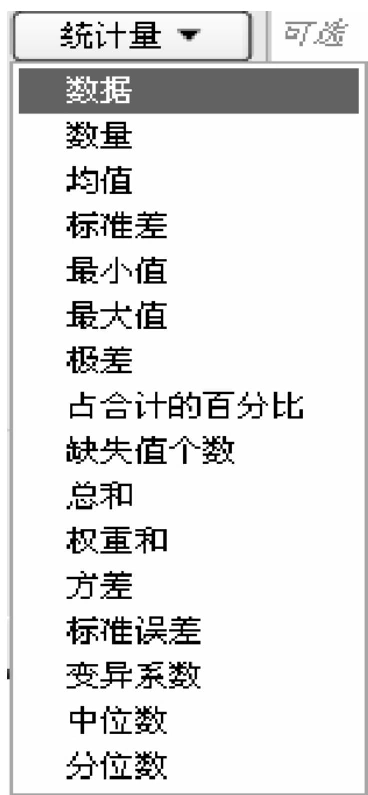


图3-22　Y 轴统计量

（3）设置Y 统计量

如果需要查看不同年龄段的体重情况，在【类别，X，水平】列表框中选择【年龄】选项，在【统计量】下拉列表框中选择【体重】选项，从可供选择的统计量下拉列表中选择均值。单击【确定】按钮，结果如图3-23所示。

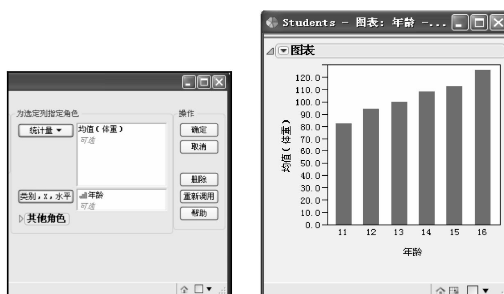


图3-23　年龄-体重图

（4）变换图表类型

JMP可以非常方便地转换图表类型，在图3-23的【图表】下拉菜单中选择【Y选项】选项，如图3-24所示。



图3-24　选择【Y选项】选项

选择不同类型生成不同的图表，如图3-25所示。

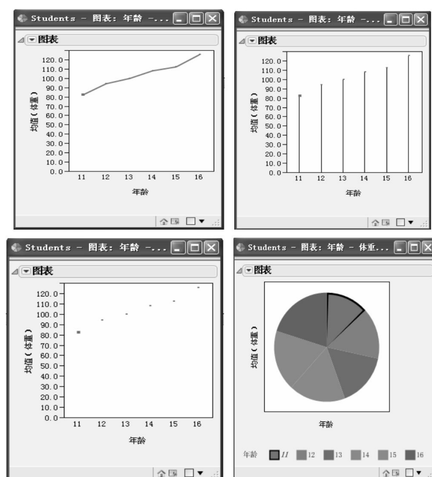


图3-25　生成不同的图表

（5）增加X 与Y 的变量数

如果想知道体重除了与年龄的关系，最好分开性别查看。在【类别，X，水平】列表框中选择【年龄】和【性别】选项作为X轴的变量，单击【确定】按钮，操作及结果如图3-26所示。

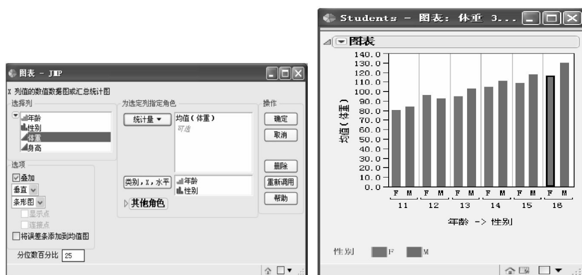


图3-26　设置两个X 变量及结果

如果不仅想看体重的变化，还要关注身高，则此时有两个X　和Y 。即增加身高的均值统计量，在【统计量】列表框中选择体重和身高的均值。单击【确定】按钮，操作及结果如图3-27所示。

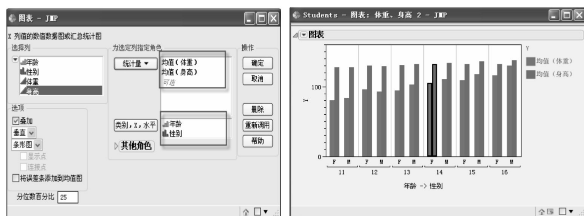


图3-27 操作及结果

（6）设置叠加图

JMP还可以轻松地设置叠加图，以方便查看分组变量。选择【图形】|【叠加图】选项，将Y 设置为身高和体重选项，X 设置为年龄，分组变量设置为性别，如图3-28所示。

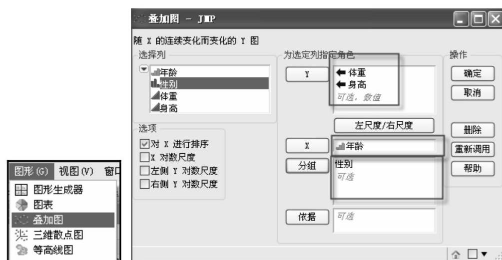


图3-28 设置叠加图

单击【确定】按钮，显示的叠加图如图3-29所示。

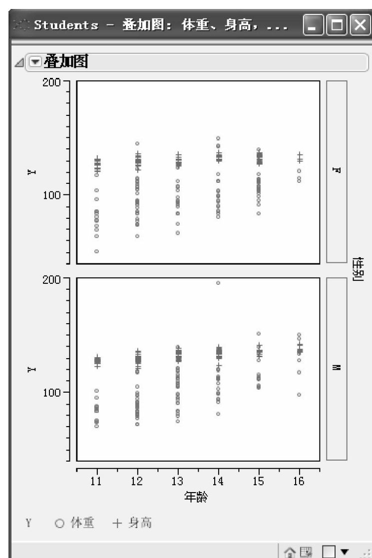


图3-29　叠加图

在此种情况下身高和体重使用同一个Y轴，但二者不是一个单位，应该分开使用两个Y坐标轴。可以在设置图形时单击【左尺度/右尺度】按钮如图3-30所示。



图3-30　【左尺度/右尺度】按钮

单击【确定】按钮，显示双坐标轴叠加图，如图3-31所示。

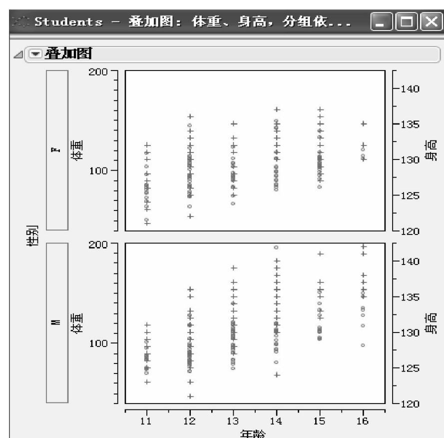


图3-31　双坐标轴叠加图

注意叠加图不接受非数值的Y轴变量。

叠加图可以绘制单个X列，以及一个或多个Y数值，在一条公共的X轴上每个Y变量的曲线可以显示为单独的图形。通过修改极差和针选项、颜色、对数轴及网格线可以修改图形，具有两种不同尺度的曲线可以在同一个图形中叠加，并外加一条右侧的轴。

3.2　图个明白——树图

树图是在有许多水平组中观测数据模式的图形，在直方图不能非常直观表现的情况下更为有用。它可以通过颜色直观地表达不同水平数，也可以称其为“热力图”。

为了更好地说明树图的作用，我们模拟52个城市的环境数据来说明，其中包括PM2.5（可吸入颗粒物）、臭氧含量和一氧化碳等指标。打开“城市.Jmp”文件，数据如图3-32所示。

	城市	X	Y	州
1	ALBANY	0.28075	0.12726	NY
2	ALBUQUERQUE	-0.15325	-0.02943	NM
3	ATLANTA	0.1652	-0.05151	GA
4	ATLANTIC CITY	0.2851	0.06908	NJ
5	BALTIMORE	0.25673	0.0617	MD
6	BOSTON	0.31556	0.13072	MA
7	BURLINGTON	0.27994	0.15985	VT
8	CHARLESTON	0.2319	0.05912	SC
9	CHARLOTTE	0.21194	-0.01885	NC
10	CHEYENNE	-0.11735	0.07356	WY
11	CHICAGO	0.10625	0.08554	IL
12	CINCINNATI	0.15238	0.04145	OH
13	CLEVELAND	0.18372	0.08804	OH
14	DENVER	-0.12191	0.04958	CO
15	DES MOINES	0.02902	0.07626	IA
16	DETROIT	0.16432	0.09998	MI
17	DUBUQUE	0.0664	0.09375	IA
18	GALVESTON-T.C.	0.01605	-0.13907	TEX
19	HARRISBURG	0.24979	0.07768	PA
20	HARTFORD	0.29779	0.11555	CT
21	HOUSTON	0.00738	-0.13123	TEX
22	HUNTINGTON	0.18168	0.03321	WV
23	INDIANAPOLIS	0.12906	0.05063	IN

图3-32　城市空气模拟数据

操作步骤如下。

（1）简单树图

树图中唯一需要的变量是类别变量，指定该变量后JMP即可产生一个树图。单击JMP菜单中的【图形】|【矩形树图】选项，指定“城市”为类别变量，如图3-33所示。



图3-33　设置矩形树图

单击【确定】按钮，生成的矩形树图如图3-34所示。



图3-34　生成的矩形树图

图中正方形的颜色来自旋转调色板，按字母顺序排列，并且按照每个组出现的次数确定大小。大多数矩形面积相等，因为每个城市只有一个数据点。Portland城有两次测量，因此其区域大小是其他城市的两倍。

（2）双变量树图

图3-34只按照城市分类，如果分类变量设置为“城市”，大小变量设置为“人口”，设置操 作及其生成的双变量树图如图3-35所示。



图3-35　设置操作及其生成的双变量树图

在图中可以清楚地看到洛杉矶和纽约的人口最多，此时的颜色变量还不能说明空气的质量好坏，下面通过一氧化碳和PM2.5等指标进

行标示。

（3）增加颜色变量

在【矩形树图】窗口中的【颜色】文本框中设置，也可以在生成的矩形树图中设置，如图3-36所示。

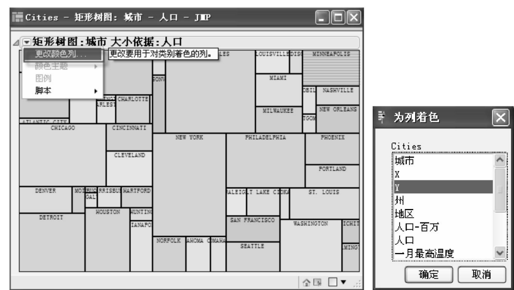
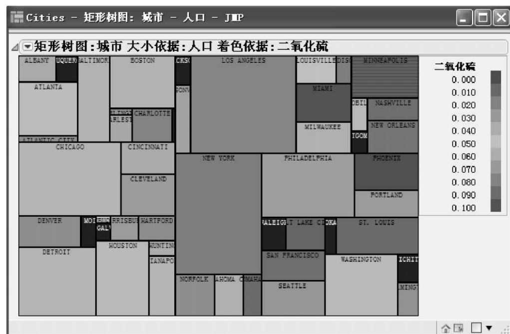


图3-36　增加颜色变量

单击窗口左上角的三角形，选择【更改颜色列】选项，显示【为列着色】对话框，分别选择【一氧化碳】、【二氧化硫】、【臭氧】和“PM200”选项，如图3-36所示。

增加颜色变量后生成的矩形树图如3-37所示。


Cities - 矩形树图: 城市 - 人口 - JMP
矩形树图:城市 大小依据:人口 着色依据:一氧化碳
一氧化碳
2.500
5.000
7.500
10.000
12.500
15.000
17.500
20.000
22.500
25.000
ALBANY
ATLANTA
BOSTON
CHARLOTTE
LOS ANGELES
MINNEAPOLIS
MIAMI
NASHVILLE
MILWAUKEE
NEW ORLEANS
CHICAGO
CINCINNATI
NEW YORK
PHILADELPHIA
PHOENIX
CLEVELAND
PORTLAND
DENVER
HARTFORD
ST. LOUIS
DETROIT
HOUSTON
SAN FRANCISCO
WASHINGTON
NORFOLK
SEATTLE


Cities - 矩形树图: 城市 - 人口 - JMP
矩形树图:城市 大小依据:人口 着色依据:二氧化硫
二氧化硫
0.000
0.010
0.020
0.030
0.040
0.050
0.060
0.070
0.080
0.090
0.100
ALBANY
ATLANTA
BOSTON
CHARLOTTE
LOS ANGELES
MINNEAPOLIS
MIAMI
NASHVILLE
MILWAUKEE
NEW ORLEANS
CHICAGO
CINCINNATI
NEW YORK
PHILADELPHIA
PHOENIX
CLEVELAND
PORTLAND
DENVER
HARTFORD
ST. LOUIS
DETROIT
HOUSTON
SAN FRANCISCO
WASHINGTON
NORFOLK
SEATTLE

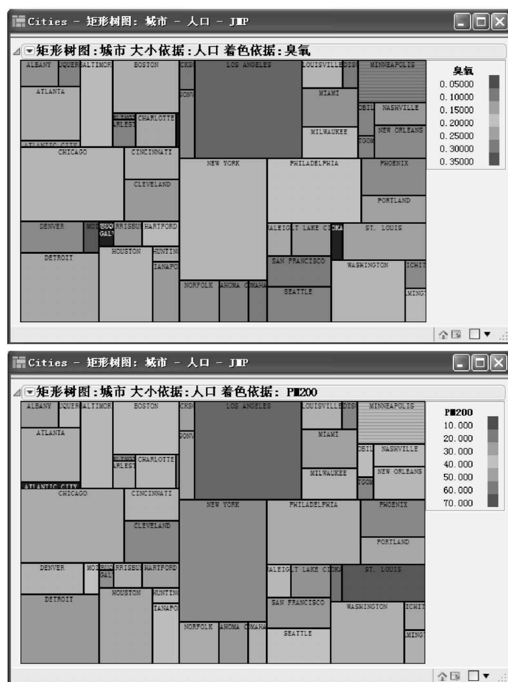


图3-37　着色之后生成的矩形树图

可以非常清晰地看出各城市的空气的污染程度，图中的黑色部分表示没有该数据，如PM2.5数据图中的“ATLANTIC　CITY”没有该测量数据。

也可以修改或自定义，单击图形中的小三角选择【颜色主题】选项后修改，如图3-38所示。

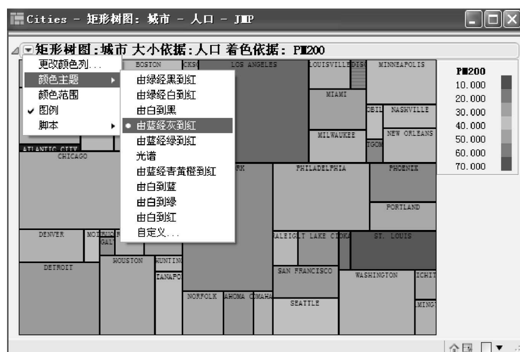


图3-38　修改颜色主题

Miss Ju：“这些图形看起来好漂亮，也很直观。”

Mr Shu：“是的，图形的优势在于直观清晰，我们列举的都是一些较为常用的图形。3.1节和本节的内容比较简单，你一定觉得很轻松吧？”

Miss Ju：“听你的口气，好像后面很难吧？”

Mr Shu：“也不难，但是作为统计类的软件，JMP和Minitab还有一种比较专业的图形，即SPC（Statistical Process Control，统计过程控制）。这是一种图形分析工具，用于判定制程是否处于统计控制状态控制范围内。”

Miss Ju：“真的不难吗？”

Mr Shu：“作为数据分析新人，我们至少应具备迎难而上的心态。SPC控制图其实也并不复杂，只是在我们常用的时间序列图上加上两条控制线，而每条控制线的计算方法不同而已。”

Miss Ju：“我们下面是不是开始学习SPC图？”

Mr Shu：“没错，JMP和Minitab的SPC图很类似，我们以

JMP例子为操作平台学习。”

3.3 图个明白——SPC图

SPC图是实施SPC的重要步骤和工具，目前在实际中大量运用的是基于Shewhart原理的传统控制图。但控制图不仅限于此，近年来又逐步发展了一些先进的控制工具。例如，对小波动进行监控的EWMA和CUSUM控制图、对小批量多品种生产过程进行控制的比例控制图和目标控制图，以及对多重质量特性进行控制的控制图。

本节重点介绍基于Shewhart原理的传统控制图，可以简单分为两种，一是计量型控制图，包括I-MR（单值移动极差图）、Xbar-R（均值极差图）和Xbar-s（均值标准差图）；二是计数型控制图，包括P（用于可变样本量的不合格品率）、NP（用于固定样本量的不合格品数）、U（用于可变样本量的单位缺陷数）和C（用于固定样本量的缺陷数）。

I-MR属于单值图，Xbar-R和Xbar-s属于子组图。单值图和子组图也称为“计量型控制图”或“变量控制图”，P、NP、U和C属于计数型数据控制图或者称为“属性控制图”。

1．SPC——单值移动控制图

（1）打开数据

打开“涂料.jmp”文件，其中记录了8组涂料共32个样本的重量，如图3-39所示。

	号码	样本	重量
1	1	1	18.5
2	2	1	21.2
3	3	1	19.4
4	4	1	16.5
5	5	2	17.9
6	6	2	19
7	7	2	20.3
8	8	2	21.2
9	9	3	19.6
10	10	3	19.8
11	11	3	20.4

图3-39　涂料重量数据表

（2）设置单值极差图

在JMP中选择【图形】|【控制图】|【单值极差】选项，显示【控制图】对话框。在【选择列】下拉列表框中选择【重量】选项，如图3-40所示。

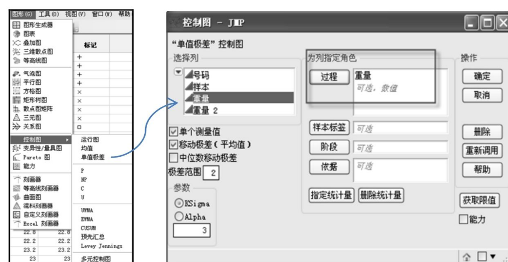


图3-40　设置单值极差图

在【过程】文本框中选择【重量】选项，单击【确定】按钮，生成的单值极差图如图3-41所示。

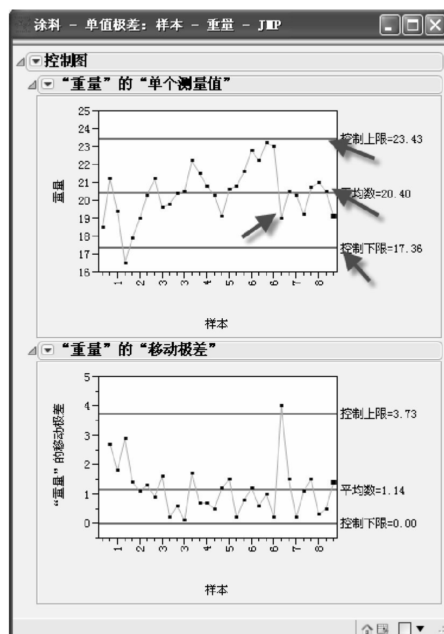


图3-41　生成的单值极差图

此控制图中有两个图表，分别是单值图和极差图，每个图有4条线。

单值图（I图）中的线条说明如下。

·折线：重量数据的单值线图。

·均值线：32组重量数据的均值线，即32组数据的算术平均值。

·上限线：$\overline{X}+3\sigma$，即均值与3倍标准差的和，只是标准差的计算与总体标准差σ的计算方法不一样。

·下限线：$\overline{X}-3\sigma$，即均值与3倍标准差之差，只是标准差的计算与总体标准差σ的计算方法不一样。

移动极差图（MR图）中的线条说明如下。

移动极差一般指一定数量数据中的最大值与最小值差值，如本例中值两个数量中最大值与最小值的差值。如果需要，也可以将计算范围设置为3和4等。

·折线：极差值的单值线图。

·均值线：31组极差数据的均值线，即31组数据的算术平均值$\overline{MR}$。

·上限线：$D4\times\overline{MR}$ 即极差与$D4$乘积，$D4$需要查表获取。

·下限线：$D3\times\overline{MR}$ 即极差与$D3$乘积，$D3$需要查表获取。如果下限值出现小于0的情况，则取0为最低限。

2．SPC——子组控制图

（1）设置子组均值极差图

前面完成的是单值极差图，如果数据中带有子组，与单值图不同的是一个子组的均值为一个单值。

选择图表类型如图3-42所示。

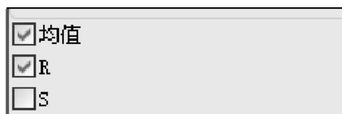


图3-42　选择图表类型

默认情况下为均值和R移动极差，即图3-41所示的图表类型；另外还有S图供选择。极差 范围默认值为2，可以根据需要设置。

设置控制图参数如图3-43所示。

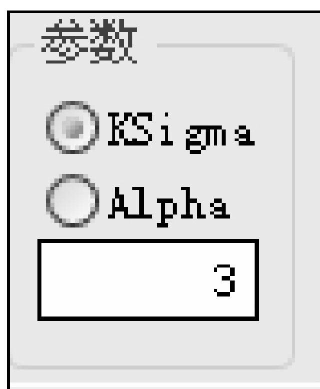


图3-43　设置控制图参数

·【K Sigma】单选按钮

选择后允许以样本标准误差的倍数计算控制限设置，它将k 个样本标准误差处的控制限指定为大于和小于期望值（显示为中心线）。要指定k ，即Sigma数，则选择【K Sigma】单选按钮并将正的k 值输入到文本框中。通常情况下k 为3，表示3个标准差。

·【Alpha】单选按钮

以单个子组统计量超出其控制限的概率，α指定控制限（也称为“概率极限”），假定流程在控制内。α的合理选项为0.01或0.001。要指定Alpha，则选择【Alpha】单选按钮并输入需要的概率。

样本标签和阶段用于对比分析不同样本数和改善前后的控制情况，如图3-44所示。

样本标签 可选
阶段 可选
依据 可选
指定统计量 删除统计量

图3-44　设置样本标签机阶段

将“样本”列输入至【样本标签】文本框，单击【确定】按钮，显示的SPC图如图3-45所示。

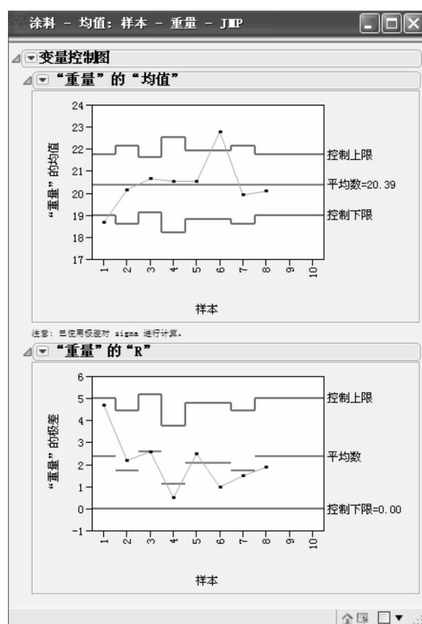


图3-45　不同样本数量SPC图

此时的SPC图成为带子组的均值极差图，与单值的均值极差图不一样的是由单值图的32个点变为8个点。因为是8组数据，所以每组数据有一个均值。每组数据的均值形成的数据构成均值极差图，其他计算方法与单值均值极差类似。

（2）设置对比图形

如果32组数据为前后两次测量结果，则在数据表中增加“阶段”列，如图3-46所示。

	号码	样本	重量	阶段
1	1	1	18.4	前
2	2	1	21.2	前
3	3	1	19.4	前
4	4	1	16.5	前
5	5	1	17.9	前
6	6	2	19	前
7	7	2	20.3	前
8	8	2	21.2	前
9	9	3	19.6	前
10	10	3	19.8	前
11	11	3	20.4	前
12	12	3	20.5	前
13	13	3	22.2	前
14	14	3	21.5	前

图3-46　修正后的数据表

在【选择列】下拉列表框中选择【阶段】选项，如图3-47所示。



图3-47　选择【阶段】选项

单击【确定】按钮，生成的单值极差图如图3-48所示。

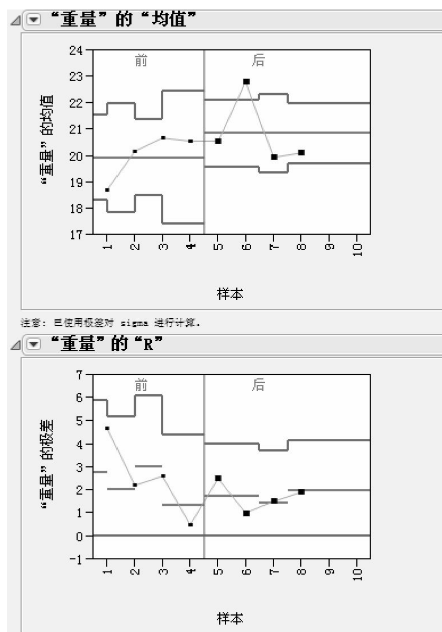


图3-48　生成的单值极差图

阶段变量可用于指定标示不同阶段或部分的列，它通常与时间对应。在此时间周期内新流程进入生产，然后经过连续的变化，为每个指定的阶段变量水平都生成一个新的Sigma、一组界限及产生的检验。

不管是计量型数据的I-MR图、$\overline{X}$-R图或计数型数据的P、U、NP和C图，均有可能有部分点位超出控制线。检测该类特殊点的原则可以参考软件的自带帮助，异常点位如图3-49所示。

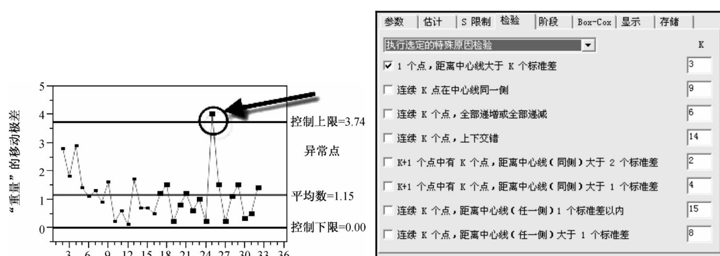


图3-49　异常点位

3．SPC——属性控制图

在前面的图类型中测量数据是过程变量，此数据经常是连续的，而且这些图是以连续理论为基础；另外一种数据类型是计数数据，其关注的变量是每个子组中缺陷数的离散计数。离散计数数据可使用属性图，因为它们是基于二项模型和Poisson模型。属性控制图在结构上类似变量控制图，只不过它们基于计数数据，而不是测量数据绘制统计图。例如，可将产品与标准进行比较并将其归类为有缺陷产品或无缺陷产品，也可以根据产品的缺陷数为产品归类。

与使用变量控制图一样，绘制过程统计量（如缺陷数）也相对于样本数量或时间，属性控制图中所标时间绘制的统计量的平均值处有一条中心线。默认情况下，属性控制图同样还会有其他两条线。即控制限制的上限和下限，它们位于中心线上下3σ处。

属性控制图的均值及上下限的计算公式请参考统计学的相关资料。

（1）缺陷品控制图（P/NP）

可以将产品与标准进行比较，并将其归类为有缺陷产品或无缺陷产品，说明如下。

·P控制图：该控制图绘制每个子组中缺陷品的比率。由于图P的每

个子组由N 个项组成，并且每项被判断为合格或不合格，因此子组中不合格项的数量最多为N。P图中 子组大小变化。

·NP控制图：该控制图绘制每个子组中缺陷品的数量。由于NP图的每个子组由Ni 个项组成，并且每项被判断为合格或不合格，因此子组i 中不合格项的最大数量是Ni 。NP图中子组大小不变。

（2）产品缺陷数控制图（C/U）

·C控制图：该控制图绘制每个子组中的缺陷数，当子组大小固定时使用C控制图。

·U控制图：该控制图绘制在每个子组中抽取的每单位样本的缺陷数，当子组大小不固定时使用U控制图。

下面以实例说明属性控制图的P/NP图。

打开“垫圈.jmp”文件，其中包含15批镀锌垫圈的缺陷计数，每批400个垫圈。对垫圈做光洁度缺陷检验，如镀面粗糙和钢材暴露。如果垫圈有光洁度缺陷，则被认为不合格或有缺陷，因此缺陷计数表示数量为400的每批中有缺陷的垫圈。文件数据如图3-50所示。

	组	组大小	次品数
1	1	400	1
2	2	400	3
3	3	400	0
4	4	400	7
5	5	400	2
6	6	400	0
7	7	400	1
8	8	400	0
9	9	400	8
10	10	400	5
11	11	400	2
12	12	400	0
13	13	400	1
14	14	400	0
15	15	400	3

图3-50　文件数据

（1）设置NP图

由于监控的是不合格品数量，而非缺陷数，因此选择P/NP图。子组大小不变，因此选择NP图。

选择【图形】|【控制图】|【NP】选项，显示【控制图】窗口，如图3-51所示。



图3-51　【控制图】窗口

在【选择列】下拉列表框中选择【组大小】选项，在【过程】文本框中输入“次品数”，在【样本标签】文本框输入“组大小”或在【样本大小】文本框中输入“400”。单击【确定】按钮，生成的NP图如下3-52所示。

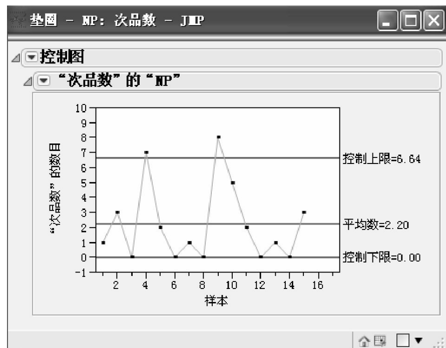


图3-52　生成的NP图

抽检的15批次样品中每批次不合格的平均数量为2.2个，控制上限为6.64，有两个批次超过控制线。

（2）设置P图

再次打开“垫圈.jmp”文件，此时的子组大小可以修改或不修改。选择【图形】|【控制图】|【P】选项，显示【控制图】窗口，如图3-53所示。



图3-53　【控制图】窗口

如果子组大小不变，在【样本大小】的【常数大小】文本框中输入“400”；如果子组大小发生变化，则在【选择列】下拉列表框中选择【组大小】选项并修改子组数，然后在【样本标签】文本框中输入“组大小”。单击【确定】按钮，生成的P图如图3-54所示。

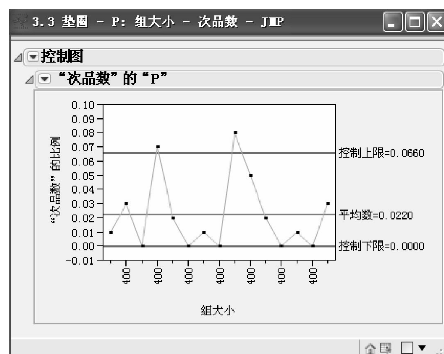


图3-54　生成的P图

尽管P图中的各点走势看上去与NP图一样，但是*Y*　轴、平均数和界限则完全不同。因为P图以缺陷产品比例为基础，而NP图以缺陷产品数量为基础。

下面以实例说明属性控制图的C/U图。

打开“支架.jmp”文件，其中记录了汽车支架每箱中的缺陷计数。每箱支架为一个检验单位，（每天）检验的箱数为子组样本大小，是可变的。文件数据表如图3-55所示。

	日期	缺陷数	单位大小
1	09FEB80	8	8
2	10FEB80	17	8
3	11FEB80	18	9
4	12FEB80	15	8
5	15FEB80	23	8
6	16FEB80	9	7
7	17FEB80	19	7
8	18FEB80	6	8

图3-55　支架缺陷数据表

由于测量的是产品的缺陷数且子组大小是变化的，所以选择U图进行监控。U图监控每个子组样本大小的支架缺陷数，上限和下限可根据检验单位数而变化。

（1）设置U图

在JMP中选择【图形】|【控制图】|【U】选项，显示【控制图】窗口。设置U图，如图3-56所示。



图3-56 设置U图

单击【确定】按钮，生成的支架U图如图3-57所示。

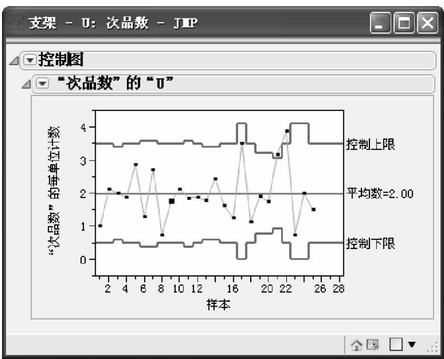


图3-57 生成的支架U图

打开“衬衣.jmp”文件，一家制衣厂以每箱10件衬衣来发货，发货前对每件衬衣进行缺陷检验。因为这家制衣厂对每件衬衣的平均缺陷数感兴趣，所以将每箱衬衣中发现的缺陷数除以10。每次检验的箱内的衬衣数都是10件，子组样本大小不变，文件数据表如图3-58所示。

	盒	缺陷数	盒装尺寸	缺陷平均数
1	1	4	10	0.4
2	2	7	10	0.7
3	3	5	10	0.5
4	4	10	10	1
5	5	3	10	0.3
6	6	2	10	0.2
7	7	0	10	0
8	8	4	10	0.4
9	9	4	10	0.4
10	10	6	10	0.6
11	11	2	10	0.2

图3-58　文件数据表

（2）设置C图

选择【图形】|【控制图】|【C】选项，显示【控制图】窗口。设置C图，如图3-59所示。



图3-59　设置C图

单击【确定】按钮，生成的衬衣C图如图3-60所示。

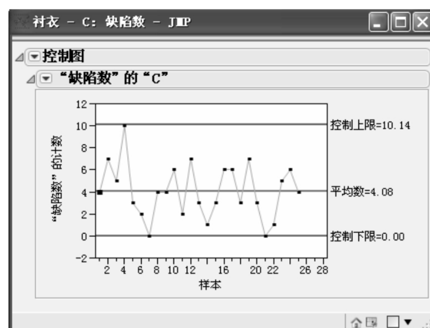


图3-60　生成的衬衣C图

C图与U图的相似之处是二者都监控整个子组中的不合格数或缺陷数，子组由一个或多个单位组成。C图要求子组大小不变，U图则反之。

控制图是一种图形分析工具，用于判定流程是否处于统计控制状态，以及监督控制内流程，它具有以下特征：

（1）每个点都代表根据质量特征度量的子组样本计算出的汇总统计量。

（2）控制图的垂直轴的标度单位与汇总统计量的单位相同。

（3）水平轴标示子组样本。

（4）当流程处于统计控制状态时图中的中心线表示汇总统计量的平均（期望）值。

（5）控制上限和控制下限分别标记为UCL和LCL，给出在流程处于统计控制状态时汇总统计量中期望的变异范围。

（6）控制限之外的点表示存在变异特殊原因。

在选择传统的控制图时可依据如图3-61所示的路径来决定。

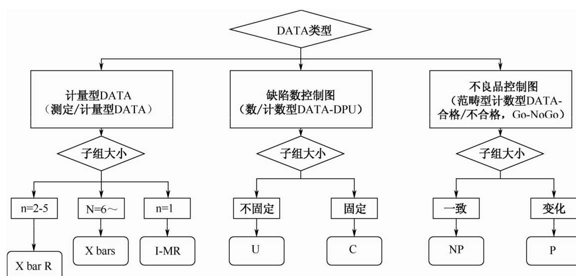


图3-61　SPC图的选择路径

Mr Shu：“第3章的内容不难吧？”

Miss Ju：“还好，但是SPC好像也不太好理解。”

Mr Shu：“当然，没有简单的学问，只有简单的学生。我们现在介绍的是入门级，所以不是特别深入。”

本章首先介绍了最常用的Pareto及饼图，说明了用Minitab和JMP制作这两种图形的差异。随后重点介绍JMP软件的操作。

JMP中有一个比较有意思的图形，即矩形树图。它是我们平时说的热力图，可以通过颜色和大小等展现数据。

最后介绍了在统计和数据分析中常用的SPC图，分别说明了连续性数据和离散型数据，并小结了如何选择控制图。

第4章
抽丝剥茧，明察秋毫
相关分析

4.1 假设检验——大胆假设，小心求证

4.2 相关与回归分析

4.3 人以类聚，物以群分——聚类分析

Mr Shu：Ju：“在第2章中我们使用Crystal Ball和LINGO进行优化分析时总是有一个目标函数要么最大，要么最小，在约束条件下求解目标值还记得吗？”

Miss Ju：“当然记得，我们要复习一下吗？”

Mr Shu：“当然不是，在求解最优解时的核心问题是目标函数。即目标与各变量之间的函数关系，这个关系是怎么来的？”

Miss Ju：“不是在解答时就已经给出来了吗？”

Mr Shu：“是的，在我们列举的案例中已给出。但在实际生产生活中这样的函数关系式需要我们自己求解，本章的重点就是探

索这些自变量与因变量之间的关系。”

Miss　Ju：“可不可以这样理解，即我们建立数学模型的重点及难点也就是寻找各变量之间的相关性，有了这个关系我们就可以建模和优化计算？”

Mr Shu：“没错，本章主要涉及的数据分析方法有假设检验、相关性分析、回归分析和聚类分析等最基本的方法，在探索这些枯燥的技术之前先讲两个比较有趣的应用案例。”

4.1　假设检验——大胆假设，小心求证

小心求证——《红楼梦》前后作者真的不是一个人吗？

2007年10月10日南京《现代快报》报道，南京林业大学汤庚国教授另辟蹊径从海棠文化角度分析《红楼梦》前80回与后40回的差异。汤教授主要从人文花卉方面分析发现，《红楼梦》前80回有16回涉及海棠，而后40回只有4回涉及海棠，以此说明前后差距明显。

汤教授从花卉角度分析，受其启发，东南大学数学系的韦博成先生从数学统计的专业角度对汤先生的发现进行数学证明。通过两个独立二项总体等价性检验并经过渐近正态公式计算，有92%的把握认为“前80回对于海棠花的关注程度大于后40回”。根据该统计方法，韦博成先生再接再厉，对于《红楼梦》中的若干重要的情景描述进行量化得到相应的数据集。有了数据集就可以进行数理统计分析，比较前80回与后40回在文风上的差异，特别是在“之乎者也”等词语的表达上。结果表明，《红楼梦》前80回与后40回在某些重要的情景描述上确实有非常显著的差异。本数据分析的目的是用数理统计的方法（具

体来说是两个独立二项总体等价性检验）来分析《红楼梦》前80回与后40回在几个重要的情景指标（包括饮食、医药、诗词、花 卉和树木的描写，这里的“描写”主要指出现的频率）的差异，并据此反映的文风来判断《红楼梦》前后两大部分差异的显著性。至于这种显著性是否能推导出作者的不同并不是本研究的目的，本数据分析研究只是数学爱好者借助自己对数学的爱好表达对生活的有趣看法。这也是数理统计爱好者的做学问态度，即大胆假设，小心求证。

研究者韦博成先生再三强调他只是从数据分析的角度指出二者的差异，尚不能说明《红楼梦》前80回与后40回作者的不同，因为“这涉及许多人文与社会方面的问题，这是数理统计方法所无能为力的。”

研究表明，饮食和花卉的显著性最高，即有充分的理由（99%）认为前80回与后40回在饮食与花卉的描述上有明显的差异，其判错的概率不到1%；医药和树木这两个指标有90%的把握认为前80回与后40回在描述上有差异。不过对于诗词的描述，并没有充分的理由发现前80回与后40回的差异。

这种研究方法是假设检验，该方法的另外一个有趣的案例是敏感的数据嗅觉——打到敌军总部。

话说林彪从带兵开始身边就有个本子，每次打完仗他就把战果记在上面，不厌其烦。不了解的人还以为他以此为乐。令人感觉得到这个人，尽管是李四光父亲的学生，23岁任军长，25岁就当军团长，但似乎有点小气。

1948年辽沈战役打响后林彪每天深夜都要进行例常的每日军情汇报，由值班参谋读出下属各个纵队、师和团用电台报告的当日战况和缴获情况。

这些几乎是重复千篇一律且枯燥无味的数据，如每支部队歼敌和俘虏多少、缴获的火炮和车辆多少，以及枪支和物资多少等。

无论战情如何紧急，人员多么疲惫，林彪依然每天坚持听军情汇报。而且要求很细，如俘虏要分清军官和士兵；缴获的枪支要统计出机枪、长枪、短枪；击毁和还能使用的汽车要分出大小和类别。

10月14日，东北野战军以迅雷不及掩耳之势仅用30小时就攻克了对手原以为可以长期坚守的锦州并全歼了守敌10余万。之后不顾疲劳，挥师北上与从沈阳出援的敌精锐廖耀湘集团20余万大军在辽西相遇，一时间形成混战。战局瞬息万变，谁胜谁负实难预料。

在大战紧急中，林彪无论多忙仍然坚持每晚必做的“功课”。一天深夜，值班参谋正在阅读下面某师上报的其下属部队的战报。其中说到他们下面的部队碰到了一个不大的遭遇战，歼敌 部分，其余逃走。与其他之前所读战报看上去并无明显异样。值班参谋就这样读着读着，林彪突然叫了一声“停！”他的眼里闪出光芒，问：“刚才念的在胡家窝棚那个战斗的缴获你们听到了吗？”

大家带着睡意的脸上出现了茫然，因为如此战斗每天都有几十起，都是差不多一模一样的枯燥数字。林彪扫视一周，见无人回答，便接连问了3句：

“为什么那里缴获的短枪与长枪的比例比其他战斗略高”？

“为什么那里缴获和击毁的小车与大车的比例比其他战斗略高”？

“为什么在那里俘虏和击毙的军官与士兵的比例比其他战斗略高”？

人们还没有来得及思索，等不及的林彪司令员大步走向挂满军用地图的墙壁指着地图上的那个点说：“我猜想，不，我断定敌人的指挥所就在这里！”

随后林彪口授命令追击从胡家窝棚逃走的那部分敌人，并坚决把他们打掉，各部队要采取分割包围的办法把失去指挥中枢后会变得混

乱的几十万敌军切成小块逐一歼灭。司令员的命令随着无线电波发向参战的各部队。

而此时的廖耀湘正庆幸自己刚刚从偶然的一场遭遇战中安全脱身并与自己的另外一支部队会合。他来不及休息就急于指令各部队尽快调整部署，为下一阶段战役做准备。可是好景不长，紧追而来的解放军迅速将其新指挥部团团围住拼命攻击，漫山遍野的解放军战士中不断有人喊着："矮胖子，白净脸。金丝眼镜湖南腔，不要放走廖耀湘！"东北野战军取得了战斗的胜利。

把对方指挥官的细节特征琢磨到如此细微并变成如此威力巨大的顺口溜，穿着满身油渍伙夫服装的廖耀湘只好从俘虏群中站出来，无奈地说"我是廖耀湘"，沮丧地举手投降。

廖耀湘对自己精心隐蔽的精悍野战司令部那么快就被发现并打掉，觉得实在不可思议，认为那是一个偶然事件，输得不甘心。当他得知林彪如何得出判断之后，这位出身黄埔军校并留学法国著名的圣西尔军校又参加过滇缅战役，在那里把日本鬼子揍得满地乱爬的新六军军长说："我服了，败在他手下不丢人。"

这两个案例说明数据分析并不神秘，古已有之。数据的收集最重要，但这是一件长期且困难的事情，利用好大数据依然需要敏锐的洞察和创新的思维。

在大数据时代来临之前，由于总体数据的参数无法获取，如全中国人成年男子的平均身高不可能一一测量，而是根据抽样学抽取一部分来测量，然后根据测量的数据来推测全国成年男子的平均身高。

假设检验是推理统计的重要分析方法，即一种利用样本来推断总体的学问。

在学习假设检验之前先来了解一下中心极限定理。

（1）无论总体服从何种分布，其多次抽样的样本均值所形成的分布往往服从正态分布。

（2）样本均值的均值与总体的均值近似。

（3）样本的标准误差（样本均值所形成分布的标准差）越小越好。

假设检验中涉及的部分名词，如表4-1所示。

表**4-1**　假设检验常用名词

范　　围	均　　值	标　准　差	比　　例
总体	μ	σ	P（大写）
样本	$\overline{X}$	S	p（小写）

假设检验通常涉及均值检验、分散检验（标准差）和比例检验等参数检验，与参数检验对应的还有中位数检验（非参数检验）。

相关性分析及回归分析是关系检验，既然要大胆假设和小心求证，就会有假设存在。例如，我们抽取一部分成年男子的平均身高数据为170　　cm。现在我们想知道推测全省成年男子的平均身高在170 cm的把握有多少，因此存在如下两种假设。

（1）零假设H0：全省成年男子的平均身高与样本数据一致，为170 cm。

（2）备择假设H1：全省成年男子的平均身高与样本数据不一样，不为170 cm。

前面我们讲过任何事情都会有风险，所以不管我们最终的结论是接受原假设还是备择假设都有可能会产生弃真或取伪错误，与这两种

错误对应的风险我们称之为“α”和“β”风险。

简单来讲，假设检验也有可能出错，只是概率比较小而已。判断假设检验结论通常有3种方法，即P 值法、置性区间法和临界值法。3种方法得出的结论相同。

通常在假设检验之前需要根据所处行业的特点指定α风险的值，即把正确的结论判断成错误的结论的风险，或者说犯弃真错误的最大容许概率值。通常来讲该值默认为0.05，即有5%的概率可能把真理说成谬论。与α 风险值相对应的为统计学上的拒绝域，1-α 为接受域。接受与拒绝指的都是零假设，即求解的Z 值或P 值在接收域则接受零假设；否则拒绝零假设选择备择假设。

如α ＝0.05，在Minitab中模拟正态分布中接收域与拒绝域如图4-1所示。

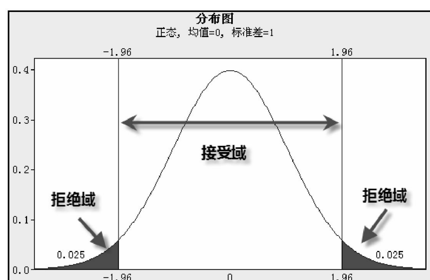


图4-1 接受域与拒绝域

假设我们已经知道全省成年男子的身高为一均值为170 cm，标准差为5 cm的正态分布，从该省份中某市的人群中抽取出来身高的数据的均值为185 cm。如果基于该数据来判断全省的成年男子的平均身高在170 cm，会得出什么样的结论？

该平均身高与全省的平均身高170 cm的距离为：

$$Z = \frac{(185-170)}{5} = 3$$

即与均值相距3个标准差，将该身高分布转换成标准正态分布的操作步骤如下。

（1）模拟分布

在Minitab中单击【图形】|【概率分布图】选项，显示【概率分布图】对话框。选择【查看概率】选项，如图4-2所示。

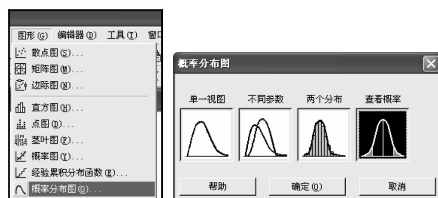


图4-2 选择【查看概率】选项

单击【确定】按钮，显示【概率分布图】对话框。单击标签，显示【阴影区域】选项卡，如图4-3所示。

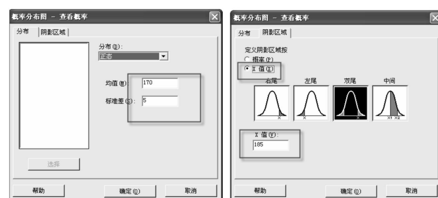


图4-3 【阴影区域】选项卡

选择【X值】单选按钮和【双尾】选项，在【X值】文本框中输入185。单击【确定】按钮，生成的正态分布模拟结果如图4-4所示。

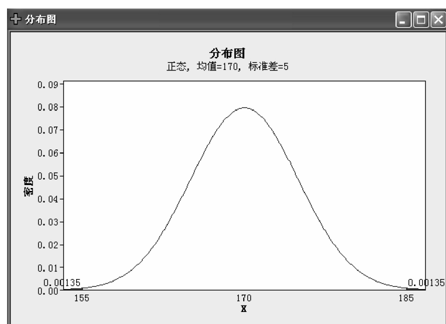


图4-4　生成的正态分布模拟结果

（2）转换分布

当$X=185$且$Z=3$时，将上面的正态分布转换成标准正态分布。

在Minitab中单击【图形】|【概率分布图】选项，显示【概率分布图】对话框。选择【查看概率】选项，均值设置为0，标准差为1，X值为3，如图4-5所示。

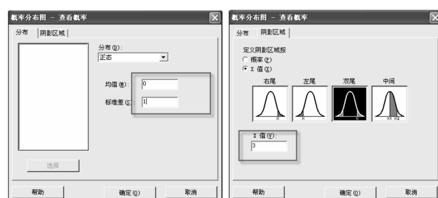


图4-5　设置转换分布

单击【确定】按钮，转换分布结果如图4-6所示。

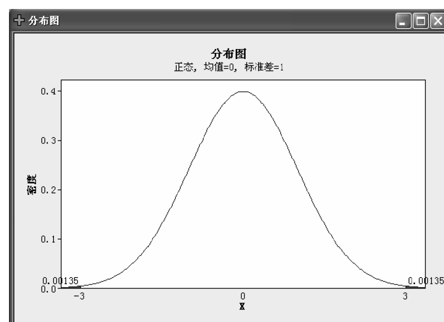


图4-6　转换分布结果

该分布转换成标准正态分布之后185的均值处在Z ＝3且P 值为0.00135的位置，与我们指定的α ＝0.05进行对比，如图4-7所示。

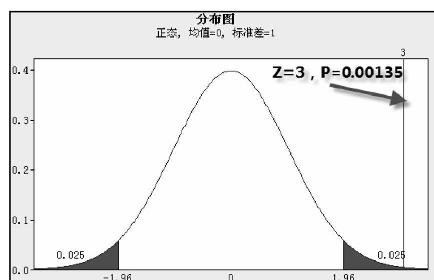


图4-7　与接收域对比

从对比看出Z ＝3与P ＝0.001 35在拒绝域内，因此我们可以应该拒绝零假设而接受备择假设。即全省成年男子的平均身高与样本数据不一样，不是170 cm。

从图解法可以看出假设检验的判断至少可以基于两个方面，即P 值与给定的α 概率判断，以及Z 值与α 对应的临界Z 值（本例中是1.96）对比得出的结论一致。

但是由于不是所有的数据都服从Z分布（正态分布），因此还有

可能是t分布或F分布等，无论何种分布都可以通过P 值与预定的α 概率进行对比。在下面的假设检验中我们均按照P 值来判断，一般与0.05来对比，即$P>0.05$，接受零假设，拒绝备择假设；$P<0.05$，接受备择假设，拒绝零假设。

4.1.1　小心求证——均值检验

1Z检验

（1）用途：检测一组样本所对应的总体的均值是否与设置的均值相等。

（2）条件：一组自变量数据，需要验证的均值和标准差。

假设H0样本所对应的总体的均值与设置的均值相同（$\mu_1=\mu_0$），H1样本所对应的总体的均值与设置的均值不相同（$\mu_1\neq\mu_0$）。

继续以身高的案例说明。

根据2005年的全国人口普查数据得知，全国18周岁以上的成年男子的平均身高为170 cm，标准差为5 cm。现在抽取某省份1 000个成年男子进行身高测量，测量数据如图4-8所示。

Minitab - 身高.MPJ - [工作表 1 ***

文件(F) 编辑(E) 数据(A) 计算(C) 统计

↓	C1	C2	C3	C4
	身高			
1	181.081			
2	194.092			
3	175.053			
4	173.224			
5	175.327			
6	192.242			
7	181.335			

图4-8　测量数据

通过统计得出该组身高数据的均值为180.43　cm，标准差为4.74 cm，是否可以说明我国的成年男子的平均身高已经超过170　cm？H0为成年男子的平均身高，与170　　cm一致；H1为成年男子的平均身高，与170 cm不一致。

（1）设置1Z检验

打开“身高.MPJ”文件，在Minitab中选择【统计】|【基本统计量】|【1Z单样本】选项，如图4-9所示。



图4-9　【1Z单样本】选项

显示【单样本Z】对话框，如图4-10所示。



图4-10　【单样本Z】对话框

在【样本所在列】下拉列表框中选择【身高】选项，标准差为总体数据的标准差，输入“5”。选择【进行假设检验】复选框，在【假设均值】文本框中输入“170”。单击【确定】按钮，结果如图4-11所示。

变量	N	平均值	标准差	平均值标准误	95% 置信区间	Z	P
身高	1000	180.430	4.736	0.158	(180.120, 180.740)	65.96	0.000

图4-11　1Z分析结果

Z =65.96远大于临界值，P =0.000＜0.005。拒绝零假设，接受备择假设，即成年男子的平均身高与170 cm不一致，平均身高是多少？继续进行假设判断。

（2）修改1Z检验设置

在【单样本Z】对话框中单击【选项】按钮，在【备择】下拉列表框中选择【大于】选项，如图4-12所示。



图4-12 【大于】选项

此时假设H0为成年男子的平均身高与170 cm一致，H1为成年男子的平均身高大于170 cm。单击【确定】按钮，分析结果如图4-13所示。

变量	N	平均值	标准差	平均值标准误	95% 下限	Z	P
身高	1000	180.074	5.017	0.158	179.814	63.71	0.000

图4-13 1Z分析结果

P ＜0.05，拒绝零假设，接受备择假设，即成年男子的平均身高大于170 cm。

如果选择【小于】选项，结果如图4-14所示。

变量	N	平均值	标准差	平均值标准误	95% 上限	Z	P
身高	1000	180.074	5.017	0.158	180.334	63.71	1.000

图4-14 1Z分析结果

从结果来看，P ＝1＝100%，即此时接受零假设，成年男子的平均身高等于170 cm。

显然不正确，这就是我们常说的风险与错误。任何的假设检验都存在风险，只是概率不一样而已。

为什么会出现这样的结果？即相对于H1（小于170　cm），选择H0（＝170 cm）会更接近正确的结果。

在本次假设的检验中我们已知总体身高的标准差为5，在更多情况下我们不知道总体标准，因此需要1t检验。与1Z检验不同，该检验不知道总体标准差，而通过t分布来判断。

1t检验

（1）用途：检测一组样本所对应的总体的均值是否与设置的均值相等。

（2）条件：一组自变量数据，需要验证的均值（总体标准差未知）。

假设H0样本所对应的总体均值与设置的均值相同，H1样本所对应的总体均值与设置的均值不相同。

继续以身高为例子说明，如果此时并不知道全省成年男子身高的标准差，需要验证全省成年男子的平均身高是否等于170 cm。与1Z相同，其假设条件H0为成年男子的平均身高与170cm一致；H1为成年男子的平均身高与170cm不一致。

（1）设置1t检验

打开“身高.MPJ”文件，在Minitab中选择【统计】|【基本统计量】|【1t单样本t】选项，如图4-15所示。



图4-15　【单样本t】选项

与1Z单样本设置相比，设置界面中已没有标准差的选项。此时需要输入假设均值，输入“170”如图4-16所示。



图4-16　设置

单击【确定】按钮，计算结果如图4-17所示。

变量	N	平均值	标准差	平均值标准误	95% 置信区间	T	P
身高	1000	180.074	5.017	0.159	(179.763, 180.385)	63.49	0.000

图4-17　计算结果

此时不是Z分布检验，而是t分布。如果采用临界值对比法，需要对比t及其临界值，直接以P 值与α 进行对比。

P ＝0.000＜0.005，拒绝零假设且接受备择假设，即成年男子的

平均身高与170 cm不一致。

（2）修正1t检验设置

将验证的均值修正为“180”，如图4-18所示。



图4-18　修正1t检验设置

此时的需要检验的假设修正H0为成年男子的平均身高与180 cm一致，H1为成年男子的平均身高与180 cm不一致。

单击【确定】按钮，分析结果如图4-19所示。

变量	N	平均值	标准差	平均值标准误	95% 置信区间	T	P
身高	1000	180.074	5.017	0.159	(179.763, 180.385)	0.47	0.641

图4-19　分析结果

P =0.641＞0.05，接受原假设，即成年男子的平均身高与180 cm一致。

可以选择3种图形进行辅助判断，单击【图形】按钮，显示【单样本t-图形】对话框，如图4-20所示。

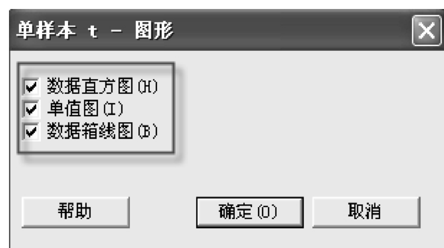


图4-20　【单样本t-图形】对话框

选择3种图形，单击【确定】按钮，生成的辅助图形如图4-21所示。

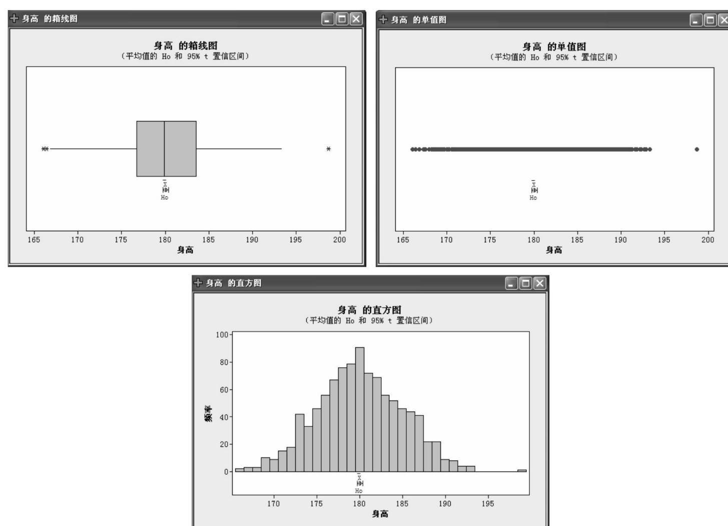


图4-21　生成的辅助图形

2t检验

（1）用途：比较两组样本所对应的总体均值是否相等。

（2）条件：两组自变量数据且两组自变量数据的方差相等。

假设H0为两组样本所对应的总体均值相等（$\mu_1 = \mu_2$），H1为两组样本所对应的总体均值不相等（$\mu_1 \neq \mu_2$）。

某超市收集了两组可乐的销售数据，分别是在超市内温度为25度以上和25度以下销售的数量。超市经理想知道超市内的温度是否对可乐的销售有影响，即需要对比在这两个温度下可乐的销售均值是否有差异。

需要检验的假设H0为25度以上和25度以下样本数据对应的可乐销售的总体均值相等（$\mu_1 = \mu_2$），H1为25度以上和25度以下样本数据对应的可乐销售的总体均值不相等（$\mu_1 \neq \mu_2$）。

（1）设置2t检验

打开“可乐销售.MPJ”文件，在Minitab中选择【统计】|【基本统计量】|【2t双样本】选项，如图4-22所示。



图4-22 【2t双样本】选项

显示【双样本t】对话框，设置有关选项，如图4-23所示。

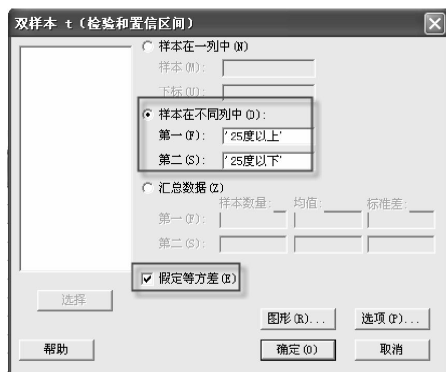


图4-23　设置2t检验

选择【样本在不同列中】单选按钮，并分别选择【'25度以上'】和【'225度以下'】两个选项，选择【假定等方差】复选框。因为2t检验的前提是两组数据的方差需相等，如果方差不等，比较均值就没有意义，这里暂定等方差。

单击【确定】按钮，分析结果如图4-24所示。

双样本 T 检验和置信区间：25度以上，25度以下

25度以上 与 25度以下 的双样本 T

	N	平均值	标准差	平均值标准误
25度以上	100	1301.8	20.9	2.1
25度以下	100	849.4	20.6	2.1

差值 = mu (25度以上) - mu (25度以下)
差值估计: 452.38
差值的 95% 置信区间: (446.59, 458.16)
差值 = 0 (与 ≠) 的 T 检验: T 值 = 154.17 P 值 = 0.000 自由度 = 198
两者都使用合并标准差 = 20.7480

图4-24　分析结果

P ＝0.000＜0.005，拒绝零假设且接受备择假设。即25度以上和25度以下样本数据对应的可乐销售的总体均值不相等，那么温度高销量好一些还是温度低销量好呢？

（2）修正2t检验设置

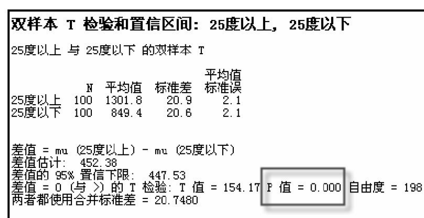
现在假设求证25度以上时可乐的销售量大于25度以下，修正2t检验，如图4-25所示。



图4-25　修正2t检验

在【备择】下拉列表框中选择【大于】选项，此时的假设变为H0为25度以上和25度以下样本数据对应的可乐销售的总体均值相等（$\mu_1 = \mu_2$），H1为25度以上对应的可乐销售的总体均值小于25度以下样本对应的可乐销售的总体均值（$\mu_1 \neq \mu_2$）。

单击【确定】按钮，分析结果如图4-26所示。



双样本 T 检验和置信区间：25度以上，25度以下

25度以上 与 25度以下 的双样本 T

	N	平均值	标准差	平均值标准误
25度以上	100	1301.8	20.9	2.1
25度以下	100	849.4	20.6	2.1

差值 = mu (25度以上) - mu (25度以下)
差值估计: 452.38
差值的 95% 置信下限: 447.53
差值 = 0 (与 >) 的 T 检验: T 值 = 154.17 P 值 = 0.000 自由度 = 198
两者都使用合并标准差 = 20.7480

图4-26　分析结果

$P = 0.000 < 0.05$，拒绝零假设，接受备择假设。即25度以上对应的可乐销售的总体均值大于　25度以下样本对应的可乐销售的总体均值。单单从这两组来看，超市内温度对可乐的销售有影响。如果经营者想增加可乐的销售，可以考虑提高超市内温度。但提高温度除了增加成本外还可能会影响其他商品的销售，因此需要权衡考虑。

在前面的2t检验中假定两组数据等方差，如果为否等方差又如

何？下面进行验证。

（3）等方差检验

在Minitab菜单中选择【统计】|【基本统计量】|【双方差】选项，显示【双方差】对话框，如图4-27所示。

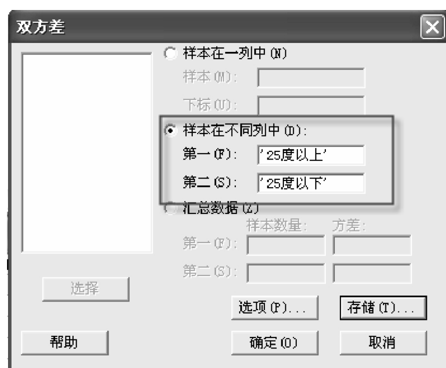


图4-27　【双方差】对话框

选择【样本在不同列中】单选按钮并分别选择【‘25度以上’】和【‘25度以下’】选项。

等方差检验的假设条件是H0为25度以上和25度以下样本数据对应的可乐销售的总体标准差相等（$\sigma_1 = \sigma_2$），H1为25度以上和25度以下样本数据对应的可乐销售的总体标准差不相等（$\sigma_1 \neq \sigma_2$）。

单击【确定】按钮，等方差检验分析结果如图4-28所示。

等方差检验：25度以上，25度以下

95% 标准差 Bonferroni 置信区间

	N	下限	标准差	上限
25度以上	100	18.0361	20.9194	24.8474
25度以下	100	17.7393	20.5751	24.4385

F 检验（正态分布）
检验统计量 = 1.03，p 值 = 0.869

Levene 检验（任何连续分布）
检验统计量 = 0.43，p 值 = 0.511

图4-28　等方差分析结果

用于判断等方差的数据有一个是F检验的P 值，该检验适用于正态数据的检验；另一个是Levene检验的P 值，该检验适用于非正态数据的检验。关于正态性的检验会在4.1.4节中说明。

F检验的P 值＝0.869和Levene检验的P 值＝0.511均大于0.05，接受零假设，即25度以上和25度以下样本数据对应的可乐销售的总体标准差相等。

基于等方差检验的前提是在前面的2t检验时选择“假定等方差”有据可依。

t-t配对检验

t-t配对检验在实际工作中经常会遇到，该检验顾名思义是成对出现，两组通常在时间上有前后关系。例如，使用两组不同的人员来进行测试考察某种减肥药的药效，一组人员吃药；一组人员不吃，很容易得到“疗效不显著”的结论，因为不同人的体重本身有很大差异。往往这种误差不是药物疗效引起的，减肥药的疗效可能淹没在总误差中。我们没有办法分辨药物引起的作用到底有多大，因此t-t配对检验时数据收集的方式是对同一组人先记录原始体重。然后记录吃药后的

体重，这样两组体重数据的对比分析才有意义。

（1）用途：比较两组样本所对应的总体的均值是否相等（同2t）。

（2）条件：两组自变量数据及其方差相等（同2t）。

假设H0为两组样本所对应的总体的均值相等（$\mu_1 = \mu_2$）；H1为两组样本所对应的总体的均值不相等（$\mu_1 \neq \mu_2$，同2t）。

某医药公司为了验证减肥药的疗效，选定10个自愿者服用一个疗程进行实验，收集的该批自愿者服药前后的数据如图4-29所示。

工作表 1 ***

↓	C1	C2	C3
	服药前	服药后	
1	90	83	
2	79	70	
3	86	80	
4	88	84	
5	92	87	
6	79	74	
7	76	79	
8	87	83	
9	102	96	
10	96	90	

图4-29　服药前后的数据

需要检验的假设H0为服药前和服药后的样本数据对应体重的总体均值相等（$\mu_1 = \mu_2$），H1 为服药前和服药后的样本数据对应的体重的总体的均值不相等（$\mu_1 \neq \mu_2$）。

（1）设置t-t检验

打开“减肥药.MPJ”文件，在Minitab中选择【统计】|【基本统计

量】|【t-t配对t】选项，显示【配对t】对话框，如图4-30所示。

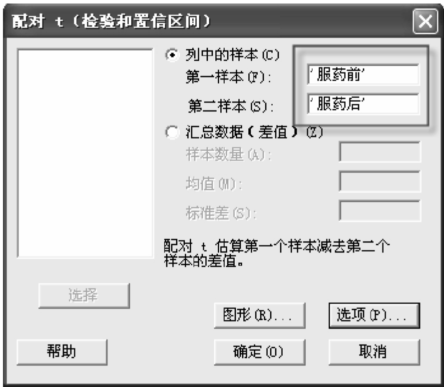


图4-30 【配对t】对话框

选中【列中的样本】单选按钮，并将服药前和服药后的体重数据分别输入【第一样本】和【第二样本】文本框中。单击【确定】按钮，分析结果如图4-31所示。

配对 T 检验和置信区间: 服药前, 服药后

服药前 - 服药后 的配对 T

	N	平均值	标准差	平均值标准误
服药前	10	87.50	8.09	2.56
服药后	10	82.60	7.52	2.38
差分	10	4.900	3.143	0.994

平均差的 95% 置信区间: (2.652, 7.148)
平均差 = 0 (与 ≠ 0) 的 T 检验: T 值 = 4.93 P 值 = 0.001

图4-31 t-t配对t检验分析结果

与2t检验一样，t-t配对检验同样按照t分布来判断；与2t检验不同，t-t配对检验在分析时Minitab中没有提示需要“假定等方差”。这是因为这两组数据来自同一个样本的不同时间段，软件默认为数据等方差。

P 值＝0.001＜0.005，拒绝零假设且接受备择假设，服药前和服药后的样本数据对应体重的总体均值不相等，即减肥药对体重有影响，效果还比较显著。

如果这两组数据用于2t检验，结果将如何？

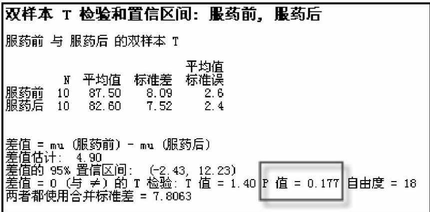
（2）设置2t检验

在Minitab中选择【统计】|【基本统计量】|【2t双样本】选项，显示【双样本t】对话框，如图4-32所示。



图4-32 【双样本t】对话框

选中【样本在不同列中】单选按钮，并将服药前和服药后的体重数据分别输入【第一】和【第二】文本框中。单击【确定】按钮，分析结果如图4-33所示。


双样本 T 检验和置信区间：服药前，服药后

服药前 与 服药后 的双样本 T

	N	平均值	标准差	平均值标准误
服药前	10	87.50	8.09	2.6
服药后	10	82.60	7.52	2.4

差值 = mu (服药前) - mu (服药后)
差值估计: 4.90
差值的 95% 置信区间: (-2.43, 12.23)
差值 = 0 (与 ≠) 的 T 检验: T 值 = 1.40 P 值 = 0.177 自由度 = 18
两者都使用合并标准差 = 7.8063

图4-33 2t双样本检验分析结果

此时$P=0.177>0.05$，接受零假设，服药前和服药后的样本数据对应体重的总体均值相等，即减肥药对体重没有影响，效果不显著。

要根据数据的收集的情况选择适当的分析方法来进行检验，否则容易犯错误。

2t检验与t-t检验的前提条件均需要等方差，只有在等方差的情况下均值检验才有意义。

单因子方差分析

2t和t-t检验都是单个因子两个水平的检验，如影响可乐销售量的因子是温度，分为25度以上和25度以下两个水平，因变量为可乐销售量。在减肥药的案例中因子为人，分为服药前和服药后两个水平，因变量为体重。二者均为验证因子水平不同是否能对因变量的均值产生影响，如果单因子的水平数增加到3个或以上时，如何验证？

（1）用途：判别两组或两组以上的样本所对应的总体均值是否相等。

（2）条件：一个因子两个或以上水平，因子在每个水平等方差。

假设H0为两组或以上样本所对应的总体的均值相等（$\mu_1=\mu_2=\mu_3$），H1为至少有两组样本所对应的总体的均值不相等（$\mu_1\neq\mu_2$或$\mu_1\neq\mu_3$或$\mu_2\neq\mu_3$或$\mu_1\neq\mu_2\neq\mu_3$）。

在化学反应中温度通常是一个非常重要的指标，其高低往往决定化学反应的剧烈程度，从而影响产品的收率。现某一化工公司想考察温度对烧碱产品收率的影响，选择了4种不同温度（120、130、140和150）试验。在同一个温度下进行10次试验，并收集每次的产品收率数据。共收集到40组数据，如图4-34所示。

↓	C1	C2	C3	C4
	120	130	140	150
1	90.4044	90.1149	91.5515	92.4128
2	89.4371	91.2046	93.6991	91.9512
3	90.2682	90.6246	91.3607	93.2927
4	89.9862	87.7925	92.1453	94.8045
5	91.7539	87.7000	92.8045	92.9664
6	92.5540	90.5473	93.5541	93.6900
7	88.9222	90.6463	93.3931	94.0300
8	91.4838	88.8531	93.8240	92.6581
9	89.1763	89.0581	94.0124	92.8434
10	89.1108	90.4475	92.3505	92.7605

图4-34　40组数据

考察温度是否会对产品收率有影响，即验证这4个温度下产品收率是否有变化需要验证的假设H0为4组温度下的产率样本所对应的总体均值相等（$\mu_1=\mu_2=\mu_3=\mu_4$），H1为4组温度下的产率样本所对应的总体均值至少有两组不相等。

（1）堆叠因子

在Minitab中方差分析需要使用堆叠格式，因此需要先进行数据堆叠。选择【数据】|【堆叠】|【列】选项，如图4-35所示。



图4-35　【列】选项

显示堆叠【列】对话框，选择C1～C4数据，单击【选择】按钮。

选择【当前工作表的列】单选按钮，输入“收率”，在【将下标存储在】文本框中输入“温度”，如图4-36所示。



图4-36　设置数据堆叠

单击【确定】按钮，数据堆叠结果如图4-37所示。

C5	C6-T
收率	温度
90.4044	120
89.4371	120
90.2682	120
89.9862	120
91.7539	120
92.5540	120
88.9222	120
91.4838	120
89.1763	120
89.1108	120
90.1149	130
91.2046	130
90.6246	130

图4-37　数据堆叠结果

（2）设置单因子方差分析

打开“化学温度.MPJ”文件，在Minitab中选择【统计】|【方差分析】|【单因子】选项，如图4-38所示。

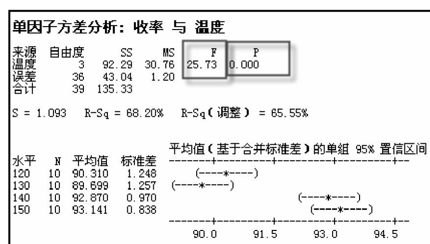


图4-38　【单因子】选项

显示【单因子方差分析】对话框，设置如图4-39所示。单击【确定】按钮，分析结果如图4-40所示。



图4-39　单因子方差设置



单因子方差分析：收率 与 温度

来源	自由度	SS	MS	F	P
温度	3	92.29	30.76	25.73	0.000
误差	36	43.04	1.20		
合计	39	135.33			

S = 1.093　R-Sq = 68.20%　R-Sq（调整）= 65.55%

水平	N	平均值	标准差
120	10	90.310	1.248
130	10	89.699	1.257
140	10	92.870	0.970
150	10	93.141	0.838

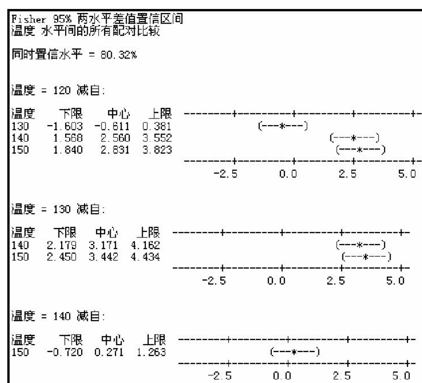
图4-40　单因子方差分析结果

简单从P 值来判断，$P=0.00<0.005$，拒绝零假设且接受备择假设，即4组温度下的产率样本所对应的总体均值至少有两组不相等，哪几组不相等？在【单因子方差分析】对话框中单击【比较】按钮，显示【单因子多重比较】对话框，如图4-41所示。



图4-41　【单因子多重比较】对话框

其中有多个选项供我们选择，没有一种选择绝对正确和精准，任何一种都有可能增加犯第1类和第2类错误的风险。我们选择【Fisher，个体误差率】复选框，并设置为5。单击【确定】按钮，比较结果如图4-42所示。



```
Fisher 95% 两水平差值置信区间
温度 水平间的所有配对比较

同时置信水平 = 80.32%

温度 = 120 减自:

温度     下限     中心    上限  --------+---------+---------+---------+-
130   -1.603   -0.611   0.381            (---*---)
140    1.568    2.560   3.552                       (---*---)
150    1.840    2.831   3.823                        (---*---)
                                --------+---------+---------+---------+-
                                     -2.5       0.0       2.5       5.0

温度 = 130 减自:

温度    下限    中心    上限  --------+---------+---------+---------+-
140   2.179   3.171   4.162                             (---*---)
150   2.450   3.442   4.434                              (---*---)
                              --------+---------+---------+---------+-
                                   -2.5       0.0       2.5       5.0

温度 = 140 减自:

温度     下限    中心    上限  --------+---------+---------+---------+-
150   -0.720   0.271   1.263              (---*---)
                               --------+---------+---------+---------+-
                                    -2.5       0.0       2.5       5.0
```

图4-42　单因子多重比较结果

从温度＝120开始分别与“130”、“140”和“150”进行对比，其结果分别为上限、中心和下限。简单判断温度＝120时产率的均值是否与其他温度时的产率相等，并观察上限与下限是否存在0，有0，则认为相等；否则不等。

因此可以判断温度＝120与温度＝130的均值相等，温度＝140与温度＝150时的均值相等。

与t-t配对检验一样，单因子方差分析均默认为等方差，必要时可以进行等方差检验。

（3）等方差检验

与2t检验和t-t检验不同，单因子方差分析已有3个以上水平，双样本的等方差检验已不能满足要求。

在Minitab中选择【统计】|【方差分析】|【等方差检验】选项，如图4-43所示。



图4-43　【等方差检验】选项

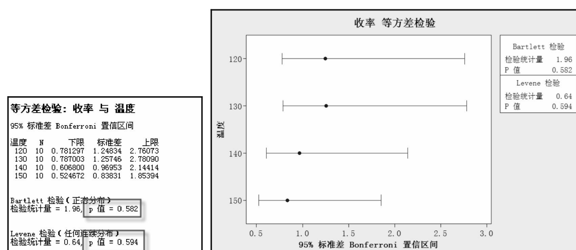
显示【等方差检验】对话框，如图4-44所示。



图4-44　【等方差检验】对话框

在Minitab中等方差检验数据也需要为堆叠格式，在【响应】文本框中输入“收率”，在【因子】文本框中输入“温度”。等方差检验的假设H0：为4组温度下的产率样本所对应的总体的方差相等（$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$），H1：为4组温度下的产率样本所对应的总体方差至少有两组不相等。

单击【确定】按钮，分析结果如图4-45所示。



等方差检验：收率 与 温度

95% 标准差 Bonferroni 置信区间

温度	N	下限	标准差	上限
120	10	0.781297	1.24834	2.76073
130	10	0.787003	1.25746	2.78090
140	10	0.606800	0.96953	2.14414
150	10	0.524672	0.83831	1.85394

Bartlett 检验（正态分布）
检验统计量 = 1.96，p 值 = 0.582

Levene 检验（任何连续分布）
检验统计量 = 0.64，p 值 = 0.594

图4-45　等方差检验分析结果

Bartlett检验用于正态分布的检验，Levene检验用于非正态数据的检验。不管是正态检验还是非正态检验，P 值均大于0.05。Bartlett检验P 值＝0.582＞0.05，接受零假设。即4组温度下的产率样本所对应的总体方差相等（$\sigma_1=\sigma_2=\sigma_3=\sigma_4$），因此步骤（2）做的单因子方差分析准确有效。

单因子方差分析结果中的MS和SS等相关术语请参考相关统计学资料。

1Z、1t、2t、tt和单因子方差分析针对的都是一个因子，如果影响因变量的是两个或多个因子且每个因子都有多个水平，则如何分析？如果是两个因子，则使用双因子方差分析；如果是多个因子，则使用多元方差分析。本书不涉及双因子方差分析等多因子方差分析，有需要的读者可以查阅相关材料。

如果自变量和因变量是计量型数据，则多因子多水平分析使用4.2节中讲解的相关与回归分析等方法。

均值检验是假设检验中常使用的参数检验，均为一个自变量的参数检验。除4.1.1节中讨论的部分均值检验除了需要等方差外，还有一个重要的条件就是正态检验，其假设如下。

（1）零假设：H0为样本数据代表的总体数据呈正态分布。

（2）备择假设：H1为样本数据代表的总体数据呈正态分布。

正态检验在Minitab中的选项为【统计】|【基本统计量】|【正态性检验】，具体操作不再叙述。

4.1.2　小心求证——比例检验

在4.1.1中我们讨论的都是检验均值，即针对计量型数据且数据均假设服从正态分布。按照数据的简单分类，还有一类重要的数据是计数型数据，如我们经常将产品分为“合格／不合格”和“良品／不良品”。抽样的结果经常是抽样100件产品，其中不合格的为4件。这种数据与连续性数据（计量型数据）有较大差异，它们是离散型变量。并且通常服从二项分布或泊松分布，比例检验就是假设检验中用来分析此类数据的方法。

比例检验顾名思义检验的不再是均值，而是比例。例如，我们有时想了解产品改善之后合格率是否显著提升或不合格率是否显著下降等，通常使用比例假设检验。

与均值检验一样，比例检验也分为一个因子、两个因子和多个因子检验，在Minitab中分别对应1P、2P和卡方检验等方法。

1P检验

1P检验也称为单“总体比例检验”，其检验的假设如下。

（1）零假设：H0为抽样的样本所代表的总体比例与设置比例相

同，即$P = P_0$。

（2）备择假设：H1为抽样的样本所代表的总体比例与设置比例不相同，即$P \neq P_0$。

随着电子产品的日益普及，小学生视力的近视成为家长和学校关心的健康问题。某学校现场随机抽取600名学生进行视力检测，其中有362名视力检测不合格，那么是否可以认为小学生的近视比例已超过六成？取$\alpha = 0.05$。

检验假设为零假设H0为抽样的学生所代表的总体近视比例与设置比例相同，即$P = 0.6$；备择假设H1为抽样的学生所代表的总体近视比例与设置比例不相同，即$P > 0.6$。

（1）设置1P检验

在Minitab中选择【统计】|【基本统计量】|【1P单比率】选项，如图4-46所示。



图4-46　【1P单比率】选项

显示【单比例（检验和置信区间）】对话框，如图4-47所示。



图4-47　【单比率（检验和置信区间】对话框

单击【选项】按钮，显示【单比率-选项】对话框，如图4-48所示。选择【汇总数据】单选按钮，在【事件数】文本框中输入“362”，在【试验数】文本框中输入“600”，并选择【进行假设检验】，在【假设比率】文本框中输入数字“0.6”。



图4-48　【单比例-选项】对话框

在【备择】下拉列表框中选择【大于】选项，注意此时是单尾检测。

单击【确定】按钮，分析结果如图4-49所示。

单比率检验和置信区间

p = 0.6 与 p > 0.6 的检验

样本	X	N	样本 p	95% 下限	精确 P 值
1	362	600	0.603333	0.569323	0.451

图4-49　单比例分析结果

概率P 值$=0.451>0.05$，因此不能拒绝零假设。即虽然小学生的抽样数据近视概率达到0.603，但并没有显著超过六成。如果此时将样本抽样人数量增加至5 000人且相应的近视人数比例也增大至3 105人，重新分析检验结果。

（2）修正参数

重新抽样后的检验设置如图4-50所示。



图4-50　重新抽样后的检验设置

在【事件数】文本框中输入“1206”，在【试验数】文本框中输入“5000”。单击【不确定】按钮，分析结果如图4-51所示。

单比率检验和置信区间

p = 0.6 与 p > 0.6 的检验

样本	X	N	样本 p	95% 下限	精确 P 值
1	3105	5000	0.621000	0.609565	0.001

图4-51　单比例检验分析结果

概率P 值$=0.001<0.05$，拒绝原假设，选择备择假设。即小学生的近视比例已显著超过六成，应该引起学校和家长的严重关注。

从600个学生中抽样有362个近视，不能判断学生的近视程度显著超过六成。而抽样5 000名学生有3 105名学生近视，我们判断学生的近视程度显著超过六成。而这二者的样本比例分别为0.603和0.62，相差不大。为什么样本比例接近，而检验结果却不相同？这主要是样本量影响分析结果。

在第2章中我们讨论概率分析时提到当二项分布的样本量够大时可以近似按照Z分布，即正态分布处理。在大样本量时，1P检验按照Z检验分析。

因此在单比例检验时需要较大的样本量支撑，单样本量较小时可能会得出错误的结论。特别是在提高客户满意度等应用上不能选取100个人，其中65人满意就得出我们的客户满意度已经超过60%的结论。

2P检验

2P检验也称为“双样本比例检验”，通常用于对比两个总体的比例。例如，了解两种生产工艺生产出来的产品合格率是否有显著差异，以及改善某种服务方式是否能显著提高客户满意度等。其检验的零假设H0为抽样的两组样本所代表的总体的比例相同，即$P_1=P_2$；备择假设H1为抽样的两组样本所代表的总体的比例不相同，即$P_1\neq P_2$。

某面条供应商为提高客户满意度选择了两种面条，即手擀面和面条机生产的面条。其中手擀面试吃1 600碗；面条机生产的面条试吃1

800碗。认为手擀面好吃的有1 212人；认为面条机生产的面条好吃的有978人，那么哪种面条更受客户欢迎？

假设检验的零假设H0为认为手擀面好吃的总体比例与认为面条机生产的面条好吃的总体比例相同，即$P_1 = P_2$。

备择假设H1为认为手擀面好吃的总体比例与认为面条机生产的面条好吃的总体比例不相同，即$P_1 \neq P_2$。

（1）设置2P检验

在Minitab中选择【统计】|【基本统计量】|【2P双比率】选项，如图4-52所示。



图4-52　【2P双比例】选项

显示【双比例（检验和置信区间）】对话框，设置如图4-53所示。



图4-53　【双比例（检验和置信区间）】对话框

选择【汇总数据】单选按钮，输入相关数据。单击【确定】按钮，分析结果如图4-54所示。

```
双比率检验和置信区间

样本      X     N    样本 p
1      1212  1600  0.757500
2       978  1800  0.543333

差值 = p (1) - p (2)
差值估计:  0.214167
差值的 95% 置信区间:  (0.183013, 0.245320)
差值 = 0(与 ≠ 0) 的检验: Z = 13.47  P 值 = 0.000

Fisher 精确检验: P 值 = 0.000
```

图4-54　2P双比例检验分析结果

概率P　值＝0.000＜0.005，拒绝零假设，选择备择假设。即认为手擀面好吃的比例比认为面条机生产的面条好吃的比例显著不同，是高还是低？需要修正备择假设。

（2）修正备择假设

如图4-55所示。

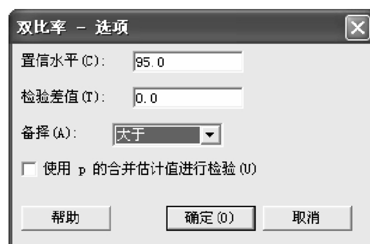


图4-55　修正备择假设双比例单尾检验

此时的假设检验修正为零假设H0认为手擀面好吃的总体比例与认为面条机生产的面条好吃的总体比例相同，即$P_1=P_2$；备择假设H1认为手擀面好吃的总体比例与认为面条机生产的面条好吃的总体比例不相同，即$P_1>P_2$。

单击【确定】按钮，分析结果如图4-56所示。

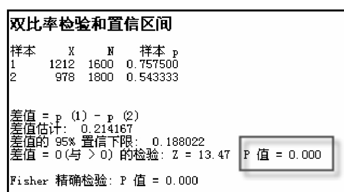


图4-56　分析结果

概率P　值$=0.000<0.005$，拒绝零假设且选择备择假设，即认为手擀面好吃的比例显著高于认为面条机生产的面条好吃的比例。

列联表与卡方检验

前面我们讨论了单总体和双总体的比例检验，当总体个数增加至3个以上时需要使用卡方检验。显然卡方检验通过卡方分布来进行检验，与IP和2P检验不同的是在Minitab中分析时卡方检验需要使用列联表格式。这是Minitab操作中常用的一个数据格式，当数据以属性或准

则分类时需要使用该格式。例如，统计各省份的男女人数就可以使用列联表格式，如表4-2所示。

表4-2　列联表格式示例

省　　份	男	女
北京	X	X
上海	X	X
广东	X	X
浙江	X	X

列联表通常使用在因子之间的卡方检验，卡方检验也称为“独立性检验”。例如，可乐的销售情况是否与其摆设位置的高或低有关，即验证因子之间是否存在相关性。下面继续以可乐的销售为例来说明卡方检验。

独立性需要需要的假设的零假设H0为因子A与因子B相互独立，备择假设H1为因子A与因子B不相互独立。

某大型连锁超市需要验证可乐的销售情况是否与其陈列方式有关，为此随机抽取了300家门店的数据。他们采用A、B和C共3种陈列方式，并统计每种陈列方式销售出去的可乐，然后按照销售量大小分为高或低，抽取的数据如表4-3所示。

表4-3　抽取的数据

销售情况	陈列方式		
	A	B	C
高	22	80	58
低	48	60	32

需要验证的零假设H0为可乐销售情况与陈列方式相互独立，备择假设H1为可乐销售情况与陈列方式不相互独立。

在Minitab中的操作步骤如下。

（1）整理数据

将表4-3中的数据整理至Minitab表中，如图4-57所示。

C1-T	C2	C3	C4
销售情况	陈列方式A	陈列方式B	陈列方式C
高	22	80	58
低	48	60	32

图4-57　可乐销量与陈列方式数据

（2）设置卡方检验

在Minitab中选择【统计】|【表格】|【卡方检验】选项，如图4-58所示。

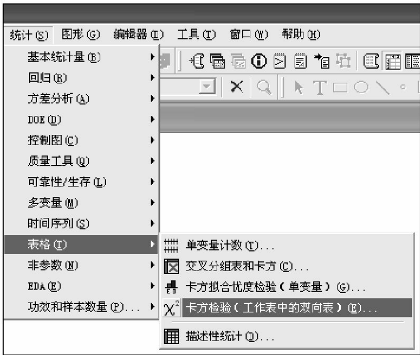


图4-58　【卡方检验】选项

显示【卡方检验】对话框，如图4-59所示。

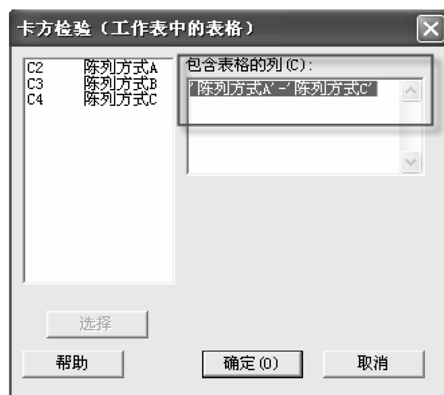


图4-59　【卡方检验】对话框

将列联表中C2～C4中的数据选择至【包含表格的列】卡拉列表框中，单击【确定】按钮，分析结果如图4-60所示。

在观测计数下方给出的是期望计数
在期望计数下方给出的是卡方贡献

	陈列方式A	陈列方式B	陈列方式C	合计
1	22	80	58	160
	37.33	74.67	48.00	
	6.298	0.381	2.083	
2	48	60	32	140
	32.67	65.33	42.00	
	7.197	0.435	2.381	
合计	70	140	90	300

卡方 = 18.776, DF = 2, P 值 = 0.000

图4-60　分析结果

概率P值$=0.000<0.05$，得到的结论是销售情况与陈列方式不具有独立性，可乐的陈列方式会影响销售情况。

如果我们的数据不是按照列联表格式排列，是否可以进行卡方检验？当然可以。如将上面案例中的数据排列成原始数据格式，如图4-61所示。

C6-T	C7	C8-T
销量	数量	方式
高	22	A
低	48	A
高	80	B
低	60	B
高	58	C
低	32	C

图4-61　原始数据格式

在Minitab中选择【统计】|【表格】|【交叉分组表和卡方】选项，显示【交叉分组表和卡方】对话框，如图4-62所示。



图4-62　【交叉分组表和卡方】选项

显示【交叉分组表和卡方】对话框，设置如图4-63所示。

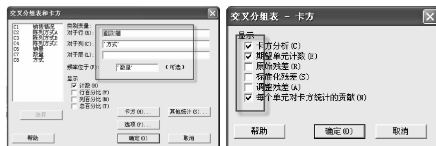


图4-63　【交叉分组表和卡方】对话框

在【对于行】文本框中输入“销量”，在【对于列】文本框中输入“方式”在，【频率位于】文本框中输入“数量”。单击【卡方】按

钮，选择【卡方分析】、【期望单元计数】和【每个单元对卡方统计的贡献】复选框。单击【确定】按钮，分析结果如图4-64所示。

汇总统计量：销量，方式

使用 数量 中的频率

行：销量　列：方式

	A	B	C	全部
低	48	60	32	140
	32.67	65.33	42.00	140.00
	7.197	0.435	2.381	*
高	22	80	58	160
	37.33	74.67	48.00	160.00
	6.298	0.381	2.083	*
全部	70	140	90	300
	70.00	140.00	90.00	300.00
	*	*	*	*

单元格内容：　计数
期望计数
对卡方的贡献

Pearson 卡方 = 18.776, DF = 2, P 值 = 0.000
似然率卡方 = 19.044, DF = 2, P 值 = 0.000

图4-64　交叉分组表卡方分析结果

概率P 值＝0.000＜0.05，得到相同的结论拒绝零假设。即可乐销售情况与陈列方式不具有独立性，3种陈列方式对销售有影响。

1P、2P和卡方检验都是基于离散变量为二项分布的情形，如果离散变量服从Poisson分布，就需要使用总体Poisson检验。在Minitab中的选项与1P和2P相似，如图4-65所示。



图4-65　Poisson率检验选项

操作方法与1P、2P和卡方检验相似，只是检验的不再是比例，而是Poisson率，操作过程不再叙述。

4.1.3　小心求证——非参数检验

Miss　Ju："说到这里，我真觉得假设检验是很神奇的工具，居然通过小小的*P* 值能看出来那么多的门道。"

Mr Shu："假设检验存在的道理就是因为我们无法去获取总体的参数，正如我们测量一个学校的学生的身高比较容易，但是测量全市、全省和全国学生的身高就很困难了。所以需要通过样本的参数来推断总体参数，*P* 值就是我们的指南针。"

Miss　Ju："可不可以这样来理解，即其实假设检验也是查找因子之间相关性的一种手段。例如，方差分析就是寻找计量型因子与计数型因子之间的相关性，而卡方分析就是为了需要计数型因子之间的相关性？"

Mr Shu："没错，我们前面讨论的几种常用假设检验方法的共同前提有二：一是观测数据要相互独立；二是数据要服从正态分

布。”

Miss Ju：“如果数据不满足上述条件怎么办？”

Mr Shu：“就要寻求新的分析方法。如果数据不独立，即数据可能完全无法表征流程的状态，此时几乎所有的理论基础都不适用。遇到这种情况的最好方法是重新审查数据的收集过程，确认是哪个不正常因素造成的影响，将不正常因素去掉后再重新收集数据。”

检验问题本身分为参数假设检验和非参数检验两大类，均值、方差或比例等检验的问题是首先数据的分布类型已明确（正态或二项）。即讨论的问题是针对参数的，这类问题称为“参数检验问题”；有些数据我们并不知道它是否服从正态或其检验的问题不是针对参数（不像参数检验中针对均值、方差和比例等），这类问题我们需要通过非参数检验来解决。

Miss Ju：“即非参数检验不需要检验类似均值和比例等字样的参数，那么它检验什么？”

Mr Shu：“非参数检验使用的方法是非参数的，它只用符号、秩和游程等几种方法分析，这是非参数检验方法的基本思想。”

Miss Ju：“符号？秩？没概念呀。万一数据也服从独立和正态，那么是不是既可以使用参数检验，也可以使用非参数检验？”

Mr Shu：“没错，数据正态的两种检验方法都是可以使用的。但统计学家早就证明了如果数据服从正态，使用参数检验要比非参数检验效果要好。但是非参数方法也有其优势，即比较简单易

懂，不需要验证很多条件就可以使用。”

Miss Ju：“简单？比参数检验简单？那我就放心了，不然我快崩溃了。”

Mr Shu：“用不着崩溃，再坚持一下，立即就可以看到胜利的曙光了。”

非参数检验——中位数符号检验法

在实际生活中我们经常需要只取两类状况的计数分析结果，如合格或不合格、满意或不满意，以及观测一组数据与其中位数的关系（分大于或小于等）。其实这类数据就是单和双比例分析问题，在单比例检验时，比例可以0～1进行假设检验，其中有一类比较特殊的比例为正好为一半。这类的比例检验用于专门检验分析比例为0.5的方法，即符号检验法。

其检验的零假设H0为抽样的两组样本所代表的总体比例相同且为0.5，即$P_1 = P_2 = 0.5$；备择假设H1为抽样的两组样本所代表的总体比例不相同，即$P_1 \neq P_2$。

在了解符号检验法之前让我们来回顾一下比例检验的案例并设置检验比例的值为0.5。

某大型电商想了解用户对目前的主流品牌手机的偏好，目前市面上主要的手机品牌升为苹果和三星。两款手机的价格比较相近，但外观和功能有所不同。现场随机抽取80个人，其中喜 欢苹果手机的人有56个人；喜欢三星手机的人有24个人。根据这个数据能判断出喜欢这两个品牌手机的人数有显著区别吗？

建立的零假设H0为喜欢苹果手机和喜欢三星手机的人数比例相同

且为0.5，即$P_1=P_2=0.5$；备择假设H1为喜欢苹果手机和喜欢三星手机的人数比例不相同，即$P_1 \neq P_2$。

操作步骤如下。

按1P检验分析，过程省略，设置和分析结果如图4-66所示。

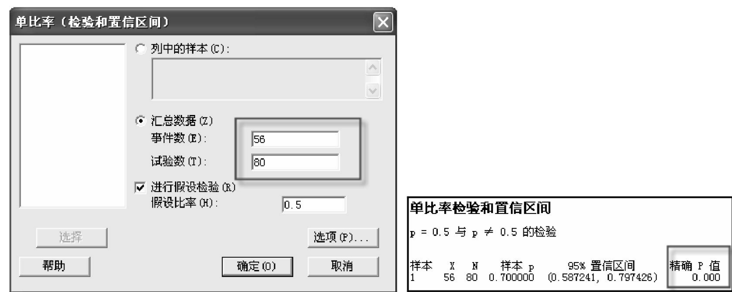


图4-66　1P设置和分析结果

概率P值$=0.000<0.05$，拒绝零假设且接受备择假设，即喜欢苹果手机的人数与喜欢三星手机的人数有显著差异。

如果数据服从正态，则均值是一组数据位置的最好代表；如果数据不服从正态或不知道是否服从正态，中位数成为最好代表。在不需要检测数据是否服从正态的情况下，我们往往关心更多的是中位数，因此对于单样本位置状况的假设变化如下。

（1）零假设H0：抽样样本所代表的总体的中位数与设置值相等，即$\eta=\eta_0$。

（2）备择假设H1：抽样样本所代表的总体的中位数与设置值不相等，即$\eta \neq \eta_0$。

这里的希腊字母η　读成“依塔”，代表中位数。如果将观测数据分别与中位数相比较，按照大小关系分为大于和小于两大类，大于记为“+”；小于记为“-”，则检验中位数的假设可以转换成检验比例，与

单比例检验相同。

前面在提到t-t配对检验时，我们已经检验过减肥药的效果，继续以其为案例说明。

（1）打开数据表

打开“减肥药.MPJ”文件，服药前后数据如图4-67所示。

C1 服药前	C2 服药后	C3 效果
90	83	7
79	70	9
86	80	6
88	84	4
92	87	5
79	74	5
76	79	-3
87	83	4
102	96	6
96	90	6

图4-67　服药前后数据

C3列设为“效果”列，保存服药前的体重减去服药后的体重数据。在Minitab计算器中设置有关选项，如图4-68所示。



图4-68　设置有关选项

在【将结果存储在变量中】文本框中输入“效果”，在【表达式】文本框中选择【‘服药前’】和【‘服药后’】两列。

从图4-67可以看出服药前后不是所有人的体重都降低，还有增长的，即服药前后的体重差值有正有负。在使用t-t配对检验时，我们通过两组数据的均值来判断减肥是否有效。通过中位数检验，如果体重差值的中位数为正数，说明体重减少的要比体重增加的人数要多，即减肥有效；反之则无效。

建立的零假设H0为两组数据的差值的中位数为0（$\eta=0$），备择假设H1为两组数据的差值的中位数大于0（$\eta>0$）。

如果将体重差值数据与0的关系分为大于和小于且分别记为“＋”和“-”，则需要检验的零假设H0为差值大于0的数据的比例为0.5，即$P_{+}=0.5$；备择假设H1为差值大于0的数据的比例大于0.5，即$P_{+}>0.5$。

（2）设置符号检验

在Minitab中选择【统计】|【非参数】|【单样本符号】选项，如图4-69所示。



图4-69　【单样本符号】选项

显示【单样本符号】对话框，如图4-70所示。



图4-70　【单样本符号】对话框

在【变量】下拉列表框中选择【效果】选项，选择【检验中位数】单选按钮后输入“0”，在【备择】下拉列表框中选择【大于】选项。单击【确定】按钮，分析结果如图4-71所示。

中位数的符号检验：效果

中位数 = 0.00000 与 > 0.00000 的符号检验

	N	下方	相等	上方	P	中位数
效果	10	1	0	9	0.0107	5.500

图4-71　分析结果

10个差值的中位数为5.5，在0的上方有9个数据；下方有1个数据。

概率P 值$=0.0107<0.05$，拒绝零假设，接受备择假设。即服药前后体重差值的中位数大于0，减肥效果显著。

样本符号检验、均值检验及比例检验的原理和应用大同小异，需求不一样，应用的检验就也不同。只不过它不需要做数据的正态性检验，因此检验功效稍低。

非参数检验——秩和检验法及M-W检验法

除了使用符号工具，非参数检验还有一个强大的工具秩，即排秩序。

如果在运动会上第2名与第3名的成绩一样，我们通常会称之为“并列亚军”。即都是第2名，第4名的人就会成为第3名。但这样会影响数据的整个排序，在检验数据的性能时可能会存在误差。因此Minitab软件在排序时通常会将两个并列第2名设置为2.5和2.5，而第4名还是会排第4，如表4-4所示。

表**4-4**　数据与秩

数据	10	11	12	12	14	14	14	15	16
秩	1	2	3.5	3.5	6	6	6	8	9

12与12应该是并列第3和第4，现在用秩表示3.5。3个14的秩序应该分为为5、6和7，现在都用6表示，15则从秩8开始排序。

排列的数据秩序与我们要检验的数据的中位数有什么关系？秩序如何来判断中位数有没有显著偏差？

如果两组数据的样本量不一样，需要来对比其的中位数是否相等。将两组数据混合在一起排出数据秩序，再将组内的秩序进行加和。中位数小的数据的秩序明显要小，而中位数大的数据秩序明显要大，因此使用秩序之和的差别是否足够大来判断两组数据的中位数是否有显著差异。

曼（Mann）和惠特尼（Whitney）在1947年将秩和检验中位数的方法进行推广，因此我们常用的双样本中位数检验方差为M-W检验，此时的中位数检验的零假设H0为抽样的两组样本所代表的总体中位数相同，即$\eta_1 = \eta_2$；备择假设H1为抽样的两组样本所代表的总体中位数不相同，即$\eta_1 \neq \eta_2$。

随着智能手机的普及，手机电池的使用时间成为用户关心的问题。某手机零配件供应商想 了解哪种品牌新手机的电池更耐用，并且选择了A和B两个品牌的新手机在同等条件下进行测试。A品牌手机测试10组数据；B品牌手机测试8组数据，测试的结果如表4-5所示。

表**4-5**　测试结果

A组	48	36	44	42	39	47	40	45	46	48
B组	38	44	36	39	37	40	41	40		

现供应商想知道这两个品牌的手机电池续航能力是否有显著差异？

建立的零假设H0为抽样的两组手机电池样本所代表的总体中位数相同，即$\eta_1 = \eta_2$；备择假设H1为抽样的的两组手机电池样本所代表的总体中位数不相同，即$\eta_1 \neq \eta_2$。

（1）整理数据

将表4-2中的数据整理至Minitab表格中并进行堆叠，如图4-72所示。

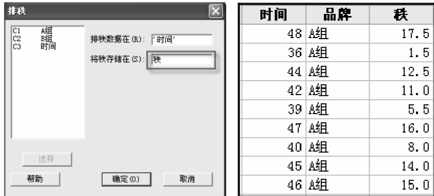


C1 A组	C2 B组	C3 时间	C4-T 品牌
48	38	48	A组
36	44	36	A组
44	36	44	A组
42	39	42	A组
39	37	39	A组
47	40	47	A组
40	41	40	A组
45	40	45	A组
46		46	A组
48		48	A组
		38	B组
		44	B组
		36	B组
		39	B组
		37	B组
		40	B组
		41	B组
		40	B组

图4-72　数据整理及堆叠

（2）数据排秩

对已堆叠好的数据进行排秩，在Minitab中选择【数据】|【排秩】选项，显示【排秩】对话框，设置及结果如图4-73所示。



时间	品牌	秩
48	A组	17.5
36	A组	1.5
44	A组	12.5
42	A组	11.0
39	A组	5.5
47	A组	16.0
40	A组	8.0
45	A组	14.0
46	A组	15.0

图4-73　设置及结果

（3）设置M-W检验

在Minitab中选择【统计】|【非参数】|【Mann-　　Whitney】选项，如图4-74所示。

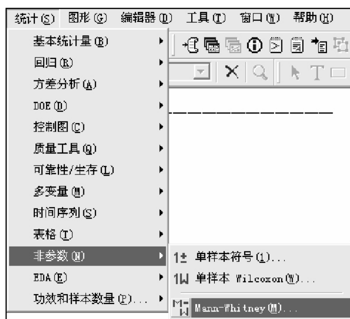


图4-74　【Mann- Whitney】选项

显示【Mann- Whitney】对话框，如图4-75所示。



图4-75　【Mann- Whitney】对话框

将A组和B组的数据分别输入至【第一样本】和【第二样本】文本框中，单击【确定】按钮，分析结果如图4-76所示。

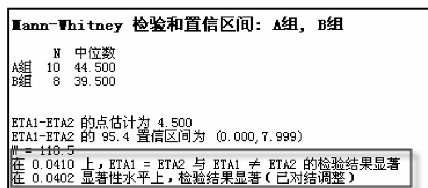


图4-76　分析结果

概率P 值＝0.0410＜0.05，拒绝零假设，接受备择假设。即抽样的两组手机电池样本所代表的总体中位数不同，两个品牌手机新电池的巡航能力有显著不同。

非参数检验——单样本Wilcoxon符号秩检验法

截止到目前我们讨论的中位数检验的方法有符号检验法和双样本的M-W检验法（秩和检验法），但符合检验法的检验功效稍低。统计学家Wilcoxon研究出了一种精密的多的“符号秩检验法”，其本质也是研究中位数或均值且检验的假设与符号检验一致，以案例说明。

如表4-6所示的简单数据组。

表**4-6**　简单数据组

22	29	71	34	32	22	42	28
21	23	36	27				

现在需要验证该批次的数据的中位数是否为24，则建立的零假设H0为该组数据的所代表的总体中位数为24，即$\eta = 24$；备择假设H1为该组数据的所代表的总体中位数为24，即$\eta \neq 24$。

将该组数据与24相减，并对绝对值分别进行排秩对比，结果如表4-7所示。

表**4-7**　对比结果

X	21	22	22	23	27	28	29	32	34	36	42	71
***X* -24**	-3	-2	-2	-1	3	4	5	8	10	12	18	47
绝对值	3	2	2	1	3	4	5	8	10	12	18	47

秩	4.5	2.5	2.5	1	4.5	6	7	8	9	10	11	12

如果此时我们使用符合检验，则将以上数据整理至Minitab中并使用单样本符号检验，过程略，设置及分析结果如图4-77所示。

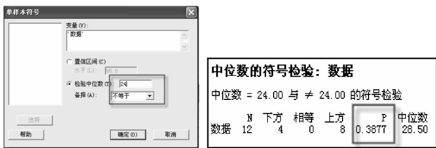


图4-77　设置及分析结果

概率P　值$=0.3877>0.05$，接受零假设，即认为该组样本所代表的总体中位数为24。

我们仔细查看数据可以看出这12个数据中小于24的有4个；大于24的有8个。并且小于24的数距离24都比较近，大于24的数据不仅个数多且距离24远。我们将数据与24进行相减，并取其绝对值。与前面不同的是对绝对值，而不是对观测值排秩。

使用Wilcoxon检验的方法的操作步骤如下：

（1）设置Wilcoxon检验

在Minitab中选择【统计】|【非参数】|【单样本Wilcoxon】选项，如图4-78所示。

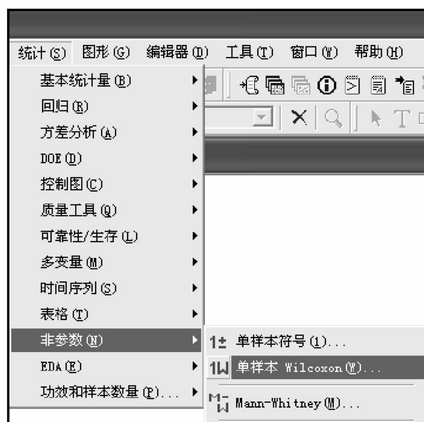


图4-78 【单样本Wilcoxon】选项

显示【单样本Wilcoxon】对话框，如图4-79所示。



图4-79 【单样本Wilcoxon】对话框

选择【检验中位数】单选按钮，输入“24”。单击【确定】按钮，分析结果如图4-80所示。

Wilcoxon 符号秩检验: 数据

中位数 = 24.00 与中位数 ≠ 24.00 的检验

	N	检验 N	Wilcoxon 统计量	P	估计中位数
数据	12	12	67.5	0.028	29.25

图4-80　分析结果

概率P 值＝0.028＜0.05，拒绝零假设，选择备择假设，即该组数据所代表的总体中位数不等于24。

相同的数据使用不同的分析方法得出的结论完全不同，Wilcoxon检验比简单的符号检验更灵敏。但与符号检验不一样的是Wilcoxon检验需要连续性数据，而符号检验则不需要。

Mr Shu："刚刚说要崩溃，现在应该感觉好多了吧？"

Miss　Ju："没有，完全没有想到假设检验会有这么多内容，有点绕不清楚了。"

Mr Shu："其实，我们前面讲解的假设检验只是一些常用的检验分析方法，归类并不复杂。我将前面讲解的内容进行了小结，如表4-8所示。

表**4-8**　常用假设检验分析方法汇总

单个 Y					
单个 X	数据类型	X 计数			X 计数
		Y 计量			Y 计数
	比较参数	均值	方差	中位数（非参数）	比例
	X 因子水平数	分析方法			
	1	1Z 检验 1t 检验	1V	1W	1P
	2	2t 检验 t-t 检验	2V	M-W/K-W	2P
	3 或以上	单因子方差（ANOVA）分析	等方差检验	M-M	卡方（Chi-Square）检验

在常用的比较分析中除了均值、中位数和比例之外还有一个常用的参数检验，即方差检验。与方差分析不同，它检验方差是否相同，而方差分析检验均值。方差检验的设置与均值检验、中位数检验和比例检验的检验假设大同小异，部分均值检验的前提是等方差。所以我们在均值 检验时需要先进行等方差检验，这在前面的例子中已经执行

过相应操作。

前面讨论的都是单个X 且X 为计数型变量的情况，如果存在多个X，则需要应用到双变量因子分析和多因子方差分析。

4.2　相关与回归分析

Mr Shu：“假设检验通过对比样本的均值、方差和中位数等参数来推断总体的情况也可以理解为寻找因子之间的相关性，通常情况下假设检验的变量（X ）为计数型数据，因变量（Y ）为计量型数据且X 变量的水平数有限，当X 的变量水平数较多时需要使用到相关和回归分析。”

Miss Ju：“变量水平变多时可不可以理解为X 变量变化为计量型数据？”

Mr Shu：“是的，可以这样理解。变量间存在的关系可以分为函数关系和相关关系两种，函数关系就是我们通常说的公式。例如，圆的面积$S = \pi y^2$ 。半径与面积之间的关系是完全确定的，即一定的直径对应一定的面积；相关关系则指变量之间存在关联，却不确定。

“相关关系是衡量变量之间的密切程度，有了相关关系变量才可以做回归分析，回归分析将变量之间的相关关系使用函数表达。有了回归分析才可以进行控制和预测，如我们前面提到的面包店的预测也是在有回归分析的基础上才可以进行的。”

Miss Ju“建立回归方程应该是建立数学模型进行预测和优化的前提，对吗？”

Mr　Shu：“是的，所以相关与回归是数据分析非常重要的环节。在学习相关与回归分析之前我们要了解相关与因果的关系，关键是理解第3变量的概念，先来看看几个案例。”

4.2.1　相关性与第三方变量

1．溺亡人数与吃冰糕的人数有关系吗？

如有统计表明游泳死亡人数越高，冰糕卖得越多。单纯从二者的数据上来看，游泳死亡人数和冰糕售出量之间呈正相关性，我们是否可以由此得出结论说吃冰糕就会增加游泳死亡风险吗？显然不可以！这两个事件都仅仅是夏天到了气温升高所导致的，吃不吃冰糕与游泳死亡风险根本没有任何因果关系。二者之间的第三方的变量就是气温，气温升高导致游泳的人增多 和冰糕的售出量的增大；游泳人多导致了溺亡人数的增加。因此我们不能得出吃冰糕会导致溺亡人数的增加；否则我们直接减少冰糕的售出量就可以减少溺亡人的数量，这显然是滑稽之谈。

2．吃海参能让人变聪明吗？

为了研究海参和聪明之间的关系，研究人员通常在一定的人群中统计他们是否平时常吃海参，挑选出常吃海参的一组和不常吃海参的一组，然后进行智商测试，并统计总体结果。看看哪一组智商平均值更高，或者直接统计吃海参频率和智商之间的相关系数。如果常吃海参的一组平均智商得分更高，那么研究人员就会得出结论，即常吃海参和智商高之间呈正相关关系。

根据这个研究，有的所谓专家声称海参吃得越多，智商就越高！为了提高智商，赶紧吃海参吧！

即便是假设常吃海参的组平均智商真的更高，并且调查对象人数真的多到了具有统计意义，“专家”的声明仍然有一个致命的逻辑缺陷。即相关性并不代表因果性！这是一个经常被人混淆，也经常被一些团体故意混淆以达到其目的。两个变量A 和B 具有相关性的原因有多种，并非只有$A \to B$ 或者$B \to A$ 这样的因果关系。一个很常见的导致相关性的可能性是A 和B 都是同样原因造成的，即$C \to A$ 并且$C \to B$。由此A 和B 也会表现出明显的相关性，但并不能说$A \to B$ 或者$B \to A$。

如我们前面提到的游泳溺亡与冰糕销量之间的关系是因为存在第3变量气温所致，只依据统计数据不足以得出因果性。想要得出因果性必须从理论上证明两个变量之间确实有因果性，并且要排除掉第3个隐含变量同时导致这两个变量的可能性。

回到海参的例子，海参和聪明之间的正相关性有可能是因为经常吃海参的家庭一般比较富裕。而富裕的家庭通常可以为孩子提供更好的教育资源，使得孩子更聪明。也可能是有一个或者多个基因同时起到使人喜欢吃海参和提升智商两种作用，如果不排除这些其他可能性，则吃海参可以导致更聪明的说法不可信。

为了纠正人们对相关性与因果关系之间的逻辑错误，可敬的科学家甚至不惜牺牲自己探索真理，着实让人佩服。

3．古德伯格与糙皮病

20世纪初，数以千计的美国南方人遭受了一种皮肤病（糙皮病）的侵袭，每年约有100 000人致命，其症状为头昏眼花、昏睡无力、伤口不愈、呕吐和严重腹泻。这种病被认为是由某种 来历不明的微生物感染所致，当时研究显示这一病症与卫生条件有联系。这一结果令全国糙皮病研究学会的医生都认为很有道理，因为南卡罗纳州没有受糙皮病干扰的人住的房子似乎都有屋内自来水及污水管道系统，卫生

条件较好。这种相关恰好验证了当时对传染病病理的理论，该理论认为由于卫生条件较差而污水处理不当，因此糙皮病通过感染者的粪便不断传播开来。

约瑟夫-古德伯格对这种解释不以为然，他对糙皮病进行了多项研究。古德伯格认为糙皮病由营养不良引起，因为美国南方地区普遍比较穷，人们吃的主要是谷物、燕麦和玉米粥这类碳水化合物高的食品，而少吃肉、鸡蛋和奶等蛋白质含量较高的事物。古德伯格认为污浊的环境和糙皮病之间的相关关系并非因果关系（就像游泳溺亡与冰糕的关系），之所以会出现相关是因为备有下水设施的家庭通常经济条件比较好。而这种经济差别也反映在他们的日常的饮食中，如是不是食用较多的动物蛋白。

但是好像有些不对劲！为什么相信古德伯格“营养缺乏”理论而不相信前面的“粪便传播”理论？毕竟两派都是坐在那里根据相关数据在推论造成糙皮病原因，古德伯格有什么理由排除“粪便处理不当”这一理论假设？古德伯格之所以理直气壮是因为他曾经亲自吃下糙皮病患者的粪便！

Mr Shu：“我们上面列举这么多例子是为了说明数据分析往往研究的是相关关系，而非因果关系。所以我们需要能准确地识别二者之间的关联，不要盲目地只从数据层面来下结论；否则可能闹笑话。”

Miss Ju：“我记住了相关关系不等于因果关系，我们要研究的是相关关系，而非因果关系。”

4．收入和支出的关系

某银行信用卡事业部想开拓市场，发展新的信用卡客户。为寻找

高消费群体，他们收集了某社区100户家庭的消费支出Y 与收入X 数据，部分数据如图4-81所示。

↓	C1	C2	C3
	收入(百元)	支出(百元)	
1	69.181	33.9461	
2	78.548	34.7722	
3	73.342	35.3924	
4	103.440	43.4419	
5	49.781	37.5716	
6	128.716	71.1788	
7	43.797	16.1815	
8	123.026	65.6007	
9	53.767	40.6944	
10	84.363	49.1335	
11	112.420	62.5844	

图4-81　部分收入与消费数据

现在银行想知道收入与支出之间是否相关？如果相关，相关程度是多少？确认相关性的操作步骤如下。

（1）制作散点图

打开“收入与支出.MPJ”文件，在Minitab中选择【图形】|【散点图】选项，显示【散点图】对话框，如图4-82所示。

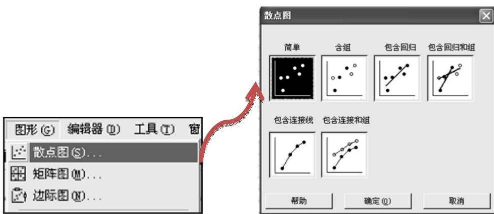


图4-82　【图形】|【散点图】选项和【散点图】对话框

选择【简单】选项，显示【散点图-简单】对话框，如图4-83所示。



图4-83　【散点图-简单】对话框

在【Y变量】中选择【'支出（百元）'】选项，在【X变量】中选择【'收入（百元）'】选项。单击【确定】按钮，生成的散点图如图4-84所示。

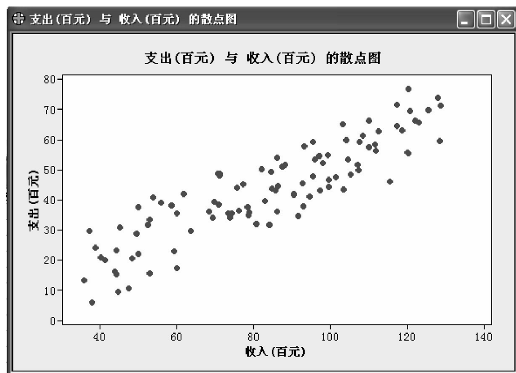


图4-84　生成的散点图

"收入"作为X轴，"支出"作为Y轴。从图中可以看出二者存在明显的相关性，相关系数是多少？

（2）计算相关系数

在Minitab中选择【统计】|【基本统计量】|【相关】选项，如图4-85所示。



图4-85 【相关】选项

显示【相关】对话框，如图4-86所示。

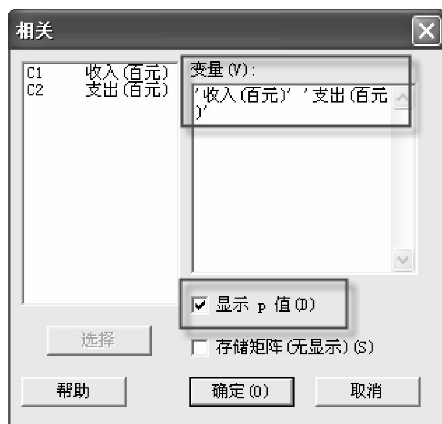


图4-86 【相关】对话框

将收入与支出两个变量输入至【变量】下拉列表框中，选择【显示P值】复选框。单击【确定】按钮，相关计算结果如图4-87所示。

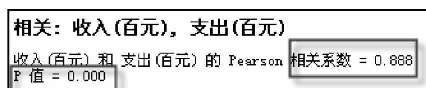


图4-87 相关计算结果

Pearson相关系数＝0.888，概率P 值＝0.000＜0.005，通过两个值可以判断这两个变量之间是否相关。相关系数的计算也应用假设检验，此时检验的零假设H0为两组变量的抽样样本所代表的总体不相关：备择假设H1为两组变量的抽样样本所代表的总体相关。

概率P 值＝0.000＜0.005，拒绝零假设，选择备择假设。即收入与支出相关，相关的密切程度通过Pearson系数查看为0.888，即强相关；相关系数＝1，为完全相关。可以理解为函数关系，如$y=2x$ 且x 与Y 始终是倍数关系，相关系数为1；相关系数为0，为完全不相关；相关系数＞0，为正相关，X 增大且Y 随之增大；相关系数＜0，为负相关，X 增大且Y 随之减少。

相关系数＝1和相关系数＝-1时的散点图如图4-88所示。

在求解收入与支出的相关性时只是两个变量之间的求解。如果变量数增加至3个以上，在Minitab中也可以快速执行多变量的相关计算。

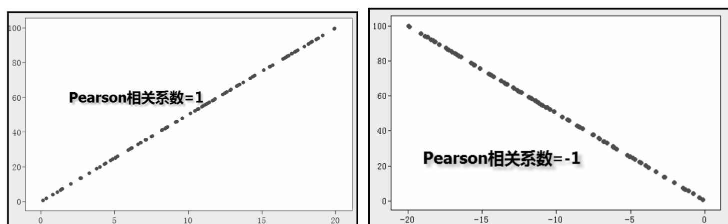


图4-88　相关系数＝1和相关系数＝-1时的散点图

（3）多变量相关系数求解

打开“溶解度.MPJ”文件，其中的数据如图4-89所示。

C1	C2	C3	C4	C5	C6
辛醇	乙醚	氯仿	苯	四氯化碳	己烷
-0.77	-1.15	-1.26	-1.89	-2.10	-2.80
-0.31	-0.57	-0.85	-1.62	-1.40	-2.10
0.25	-0.02	-0.40	-0.70	-0.82	-1.52
0.88	0.89	0.45	-0.12	-0.40	-0.70
1.56	1.20	1.05	0.62	0.40	-0.40
2.03	1.80	1.69	1.30	0.99	0.46
2.41	2.40	2.41	1.91	1.67	1.01
-0.17	-0.34	-1.60	-2.26	-2.45	-3.06
0.33	0.27	-0.96	-1.35	-1.60	-2.14
0.79	0.61	-0.27	-0.96	-0.97	-1.76
1.92	1.95	1.15	0.30	0.57	-0.46
1.39	1.00	0.28	-0.10	-0.42	-1.00
1.33	1.21	-0.69	-1.30	-1.66	-2.63
0.92	1.31	-0.89	-1.40	-2.31	-2.72
0.22	0.37	-1.92	-1.60	-2.56	-3.14

图4-89　溶解度数据

在Minitab中选择相关系数分析选项，显示的【相关】对话框如图4-90所示。



图4-90　【相关】对话框

在【变量】下拉列表框中选择C1～C6列数据，单击【确定】按钮，多元相关计算结果如图4-91所示。

相关：辛醇，乙醚，氯仿，苯，四氯化碳，己烷

	辛醇	乙醚	氯仿	苯	四氯化碳
乙醚	0.934 0.000				
氯仿	0.598 0.000	0.515 0.000			
苯	0.720 0.000	0.649 0.000	0.941 0.000		
四氯化碳	0.615 0.000	0.538 0.000	0.921 0.000	0.957 0.000	
己烷	0.605 0.000	0.549 0.000	0.872 0.000	0.909 0.000	0.950 0.000

单元格内容：Pearson 相关系数
　　　　　　P 值

图4-91　多元相关计算结果

从计算结果可以看出大部分溶解度数据强烈相关，概率P　值均为0.000。

（4）多变量散点图

此时的变量数已增至6，简单散点图已不能满足需求，可以使用矩阵图完成。在Minitab中选择【图形】|【矩阵图】选项，显示【矩阵图】对话框，如图4-92所示。

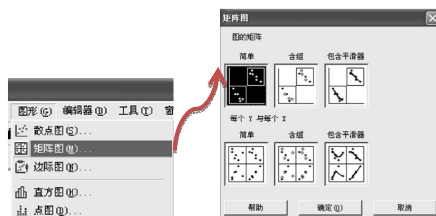


图4-92　【矩阵图】选项和【矩阵图】对话框

选择【简单】选项，显示【矩阵图-图矩阵-简单】对话框。在【图形变量】下拉列表框中选择C1～C6变量选项，如图4-93所示。单击【确定】按钮，生成的多变量矩阵图如图4-94所示。



图4-93　设置有关选项

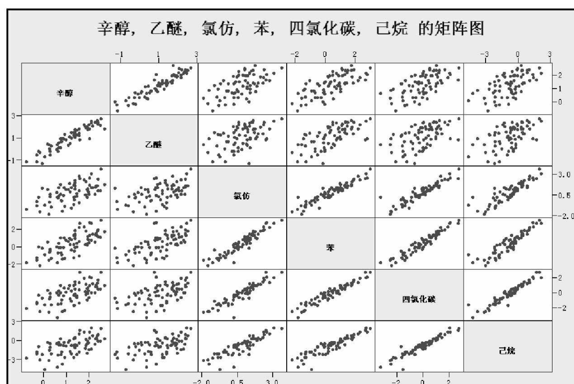


图4-94　生成的多变量矩阵图

从矩阵图也可以看出各变量之间的相关程度，相关系数高只能说明二者之间存在相关关系，而非因果关系，因为有可能还有第3变量的存在。在我们寻找根本原因时相关系数是一个参考值，是否为决定值还需要进一步深入分析数据之间的逻辑关系。

4.2.2　收入与支出关系——简单线性回归

Mr　Shu：“Miss　Ju，想必你一定知道我们为什么要分析相关

性吧？”

Miss　　Ju：“有相关性的情况下我们用来回归变量之间的关系，这样可以通过自变量来控制因变量。”

Mr　Shu：“是的，我们还一定要小心第3变量的存在。如果确实存在相关关系，则可以用来改善目标变量了。”

回归分析分类也比较多，初学者应重点掌握简单线性回归、多元线性回归、非线性回归和离散变量的Logistic回归。

回到收入与支出的相关分析案例，我们做简单线性回归。在该案例中因变量是支出，自变量为收入，操作步骤如下。

（1）打开“收入与支出.MPJ”文件，在Minitab中选择【统计】|【回归】|【回归】选项，如图4-95所示。



图4-95　【回归】选项

显示【回归】对话框，如图4-96所示。

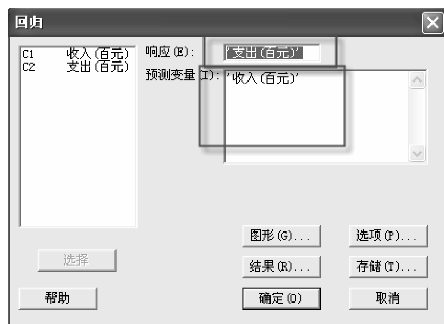


图4-96 【回归】对话框

在【响应】下拉列表框中选择【支出（百元）】选项，在【预测变量】下拉列表框中选择，【‘收入（百元）’】选项。单击【确定】按钮，结果如图4-97所示。

```
回归分析: 支出(百元) 与 收入(百元)

回归方程为
支出(百元) = - 1.66 + 0.539 收入(百元)

自变量        系数    系数标准误       T      P
常量        -1.665       2.462   -0.68  0.500
收入(百元)  0.53864     0.02814   19.14  0.000

S = 7.29006   R-Sq = 78.9%   R-Sq(调整) = 78.7%

方差分析

来源      自由度     SS     MS       F      P
回归           1  19473  19473  366.41  0.000
残差误差      98   5208     53
合计          99  24681

异常观测值

           收入(百                            拟合值             标准化
观测值       元)  支出(百元)  拟合值    标准误     残差     残差
    55       115      45.981  60.503     1.155  -14.522   -2.02R

R 表示此观测值含有大的标准化残差
```

图4-97 回归分析结果

从拟合方程上来看，支出（百元）＝-1.66＋0.539收入（百元）。即支出与收入呈正比，收入每增加10个单位，则支出会相应多出3.73的单位。

因此如果想大胆支出消费，还是要多赚钱。

在回归分析中我们再次看到方差分析，与前面假设检验中提到的

方差分析相同。假设我们用来检验一个自变量的多个水平是否会对因变量的均值产生影响，这里用方差分析的方法对回归方差做显著性检验，需要检验的零假设H0为回归方程对模型无意义，备择假设H1为回归方程对模型有意义。方差分析通过F分布来判断，并按照概率P 值来判断。

概率P 值＝0.000＜0.005，回归方程有意义。

除了方差分析的P 值，在回归分析中还有回归方程的系数的显著性检验，我们的拟合方程为：

支出（百元）＝-1.66＋0.539收入（百元）

显然0.539的系数要比-1.66的常量系数要重要，因为如果0.539这个系数变为0，则回归方程就变成Y ＝-1.66，没有任何拟合意义。因此我们在查看系数显著性分析时，如果斜率系数较为显著，则可以不用太关注常量系数。

方差分析中每个数据的详细计算可以参考其他专业书籍。

在Minitab中执行简单的线性回归还可以通过拟合功能完成，选择【统计】|【回归】|【拟 合线图】选项，如图4-98所示。



图4-98 【拟合线图】选项

其后操作过程略，在拟合线图中拟合简单线性只能完成一个自变量与因变量的拟合。如果自变量较多，拟合线性方程还是通过菜单中

的【回归】选项完成。

4.2.3　最佳口感食品配方——多元线性回归

1．改善食品口感

某食品公司最新推出来一款休闲食品，其中主要包括3种原料。为了获得最佳的配方，在上市之前该公司测试了29种配方并邀请一批美食专家对每种配方生产出来的食品进行测试评分，配方与评分数据如图4-99所示。

↓	C1 口感总分	C2 原料1	C3 原料2	C4 原料3
1	271.8	33.53	40.55	16.66
2	264.0	36.50	36.19	16.46
3	238.8	34.66	37.31	17.66
4	230.7	33.13	32.52	17.50
5	251.6	35.75	33.71	16.40
6	257.9	34.46	34.14	16.28
7	263.9	34.60	34.85	16.06
8	266.5	35.38	35.89	15.93
9	229.1	35.85	33.53	16.60
10	239.3	35.68	33.79	16.41
11	258.0	35.35	34.72	16.17
12	257.6	35.04	35.22	15.92
13	267.3	34.07	36.50	16.04

图4-99　食品配方与评分数据

现在想知道哪些原料对口感有显著影响？以提高食品的口感分数。

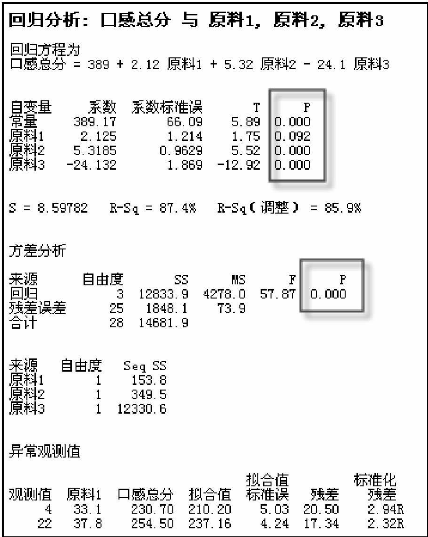
由于此时的自变量为3个且分别有29个水平，因此使用多变量回归分析，操作步骤如下。

打开“原料配比.MPJ”文件，在Minitab中打开【回归】对话框，如图4-100所示。



图4-100　【回归】对话框

在【响应】下拉列表框中输入“口感总分”，在【预测变量】下拉列表框中选择C2～C3列。单击【确定】按钮，分析结果如图4-101所示。



回归分析：口感总分 与 原料1，原料2，原料3

回归方程为
口感总分 = 389 + 2.12 原料1 + 5.32 原料2 - 24.1 原料3

自变量	系数	系数标准误	T	P
常量	389.17	66.09	5.89	0.000
原料1	2.125	1.214	1.75	0.092
原料2	5.3185	0.9629	5.52	0.000
原料3	-24.132	1.869	-12.92	0.000

S = 8.59782　R-Sq = 87.4%　R-Sq（调整）= 85.9%

方差分析

来源	自由度	SS	MS	F	P
回归	3	12833.9	4278.0	57.87	0.000
残差误差	25	1848.1	73.9		
合计	28	14681.9			

来源	自由度	Seq SS
原料1	1	153.8
原料2	1	349.5
原料3	1	12330.6

异常观测值

观测值	原料1	口感总分	拟合值	拟合值标准误	残差	标准化残差
4	33.1	230.70	210.20	5.03	20.50	2.94R
22	37.8	254.50	237.16	4.24	17.34	2.32R

图4-101　多元线性回归分析结果

从分析结果来看，拟合的方程为：

口感总分＝389＋2.12原料1＋5.32原料2-24.1原料3。即如果想提高口感的分数，应该提高原料1和原料2的比例，减少原料3的比例。

从系数的显著性分析结果来看，原料1和原料3对口感的影响比较显著，分别为正相关和负相关。

原料2的系数显著性P 值为0.092＞0.05，可能对口感的影响并不显著。但由于P 值＝0.092超出0.05也不太多，因此有待进一步考证。

除了可以通过数据来查看显著性，还可以通过残差图形来分析。在【回归】对话框中单击【图形】按钮，显示【图形】对话框。选择【四合一图形】单选按钮，单击【确定】按钮，生成的图形如图4-102所示。

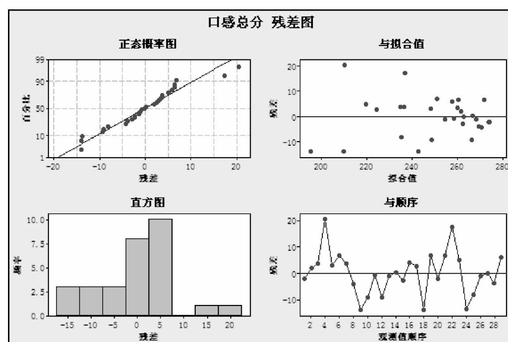


图4-102　生成的图形

此图为残差分析图，残差为观测值与拟合值的差值。如第1组观测值为271.8，拟合值为389＋2.12×33.53＋5.32×40.55-24.1×16.66＝274.3，二者的差值为271.8-274.3＝-2.5即残差。这样共有29个残差值。残差图的解读如下。

（1）右下角的残差与观测值顺序图重点考察此图中的残差值是否随机地在水平轴上下无规 则地波动，如果是，说明残差相互独立；否则呈现一定规则的图形，如全部在水平轴以上或以下。回归分析需要仔细检查，在本例中呈上下随机波动。

（2）右上角的残差与拟合值的X轴为拟合值，Y轴为残差值。重点考察残差是否保持等方差性，主要观察图形是否有“漏斗形”或“喇叭形”。如果图形有异常，说明因变量Y需要做某种变换后重新拟合。

（3）左上角的正态概率图为残差的正态概率图，主要考察残差是否符合正态分布。

（4）左下角的直方图重点考察是否有弯曲趋势，如果有，说明线性回归可能不是最佳的拟合曲线，考虑高阶拟合可能会使模型更加吻合。

2．减少氨气损失量

在氨制冷过程中的氨气量非常重要，不足会导致制冷量的不足，因此在此过程中需要尽量减少氨气的损失量。根据历史经验，氨气的损失量可能与反应过程中的气流、水温及酸浓度有关。为了验证历史经验是否正确和减少氨气的损失量，现对氨气损失量进行回归分析，收集的数据如图4-103所示。

C1	C2	C3	C4	C5
常数	气流	水温	酸浓度	损失量
1	80	11.9	85	42
1	80	11.7	88	35
1	75	9.2	90	34
1	62	8.5	87	22
1	62	6.9	87	18
1	62	7.9	87	18
1	62	8.6	93	19
1	62	8.5	93	20
1	58	7.4	87	15
1	58	2.6	80	14
1	58	2.5	89	14
1	58	1.9	88	13
1	58	2.7	82	11
1	58	3.6	93	12
1	50	2.3	89	6
1	50	2.8	86	7
1	50	3.6	72	9
1	50	3.6	79	8
1	50	4.9	80	9
1	56	4.4	82	15
1	70	5.1	91	20

图4-103　收集的数据

该数据表中因变量是损失量，自变量为气流、水温和酸浓度。打开“减少氨损失量.MPJ” 文件，对其进行多变量的线性回归。过程同上例，拟合的分析结果如图4-104所示。

```
回归分析: 损失量 与 气流, 水温, 酸浓度

回归方程为
损失量 = - 20.9 + 0.815 气流 + 0.844 水温 - 0.186 酸浓度


自变量        系数  系数标准误       T      P
常量       -20.909       7.563   -2.76  0.013
气流       0.81530     0.08351    9.76  0.000
水温        0.8440      0.2320    3.64  0.002
酸浓度    -0.18602     0.09224   -2.02  0.060


S = 1.99074   R-Sq = 96.3%   R-Sq(调整) = 95.6%


方差分析

来源          自由度       SS      MS       F      P
回归               3  1751.87  583.96  147.35  0.000
残差误差          17    67.37    3.96
合计              20  1819.24


来源      自由度   Seq SS
气流           1  1681.43
水温           1    54.32
酸浓度         1    16.12


异常观测值

                                   拟合值                标准化
观测值  气流  损失量  拟合值  标准误    残差      残差
     1  80.0  42.000  38.548   1.191   3.452     2.16R
    21  70.0  20.000  23.539   1.010  -3.539    -2.06R

R 表示此观测值含有大的标准化残差
```

图4-104　氨损失量拟合的分析结果

从拟合结果看回归总效果显著，在3个自变量系数显著性分析中

只有酸浓度的概率P值$=0.06>0.05$，该变量是否保留还需要再讨论。打开残差的四合一图形，如图4-105所示。

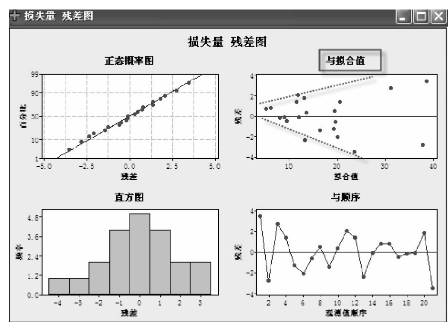


图4-105　残差的四合一图形

似乎残差与拟合值图有呈喇叭状的发散状况，随着拟合值的增大残差的上线波动幅度越来越大。这是提醒我们是否要对数据进行变换，操作步骤如下。

（1）启用宏命令

在Minitab中选择【编辑】|【命令行编辑器】选项，如图4-106所示。



图4-106 【命令行编辑器】选项

对因变量所采用的变换称为“Box-Cox变换”，其计算公式如下：

$$y^* = \begin{cases} y^{\lambda} & (\lambda \neq 0) \\ \ln y & (\lambda = 0) \end{cases}$$

求λ 值之后对其进行变换，宏指令是Boxcoxregres，其使用规则简述如下。

·数据需要存放在连续列中，如本例中的3个自变量和一个因变量都是存在连续的若干列中且左侧必须有一列全部是1的常数列。

·命令的格式为%boxcoxregres因变量列数常数列-自变量列，本例的命令格式为%boxcoxregres C5 C1-C4。

（2）输入宏命令

显示【命令行编辑器】对话框，在宏命令文本框中输入命令，如图4-107所示。

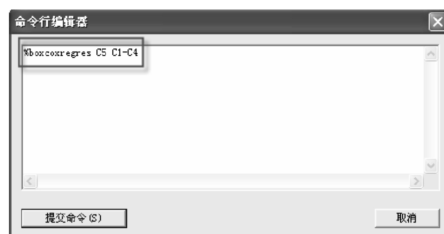


图4-107　输入宏命令

简单判断数据是否需要变换的原则如下。

·如果曲线落在水平的点虚线之下，则表示对应此范围内λ值都能满足线性回归模型的条件。

·重点考察λ值＝1时曲线是否落入点虚线下方，如果落入，则不需要变换；如果高于点虚线，则表示*y*　必须变换才能满足拟合需要。

从图中可知λ＝1时，曲线落在水平的点虚线上方，此时需要对*y*进行变换；当l＝0.5时曲线达到最低，此时需要进行的变换是取*y*　的0.5次方。

（3）变换*Y* 值

在原表中对*Y*　值进行开方变换数据，在Minitab中选择【计算】|【计算器】选项，显示的【计算器】对话框，如图4-108所示。



图4-108 【计算器】对话框

在其中执行开方操作，将开方的数据存在“开方”列中。因为本次的λ值正好是0.5，所以可以开方；如果是其他值，则需要执行指数运算。

（4）重新进行多元回归

重新对开方后的数据进行多元回归，回归后的拟合结果如图4-109所示。

回归方程为
开方 = - 0.992 + 0.0896 气流 + 0.0953 水温 - 0.0112 酸浓度

自变量	系数	系数标准误	T	P
常量	-0.9919	0.8541	-1.16	0.262
气流	0.089635	0.009431	9.50	0.000
水温	0.09525	0.02620	3.64	0.002
酸浓度	-0.01115	0.01042	-1.07	0.299

S = 0.224823   R-Sq = 96.3%   R-Sq(调整) = 95.6%

图4-109 变换后的拟合结果

3个变量系数的显著性分析中只有酸浓度不显著，但由于Y是变换之后的拟合，所以酸浓度是否显著还是需要论证原始数据后确定。残差分析图如图4-110所示。

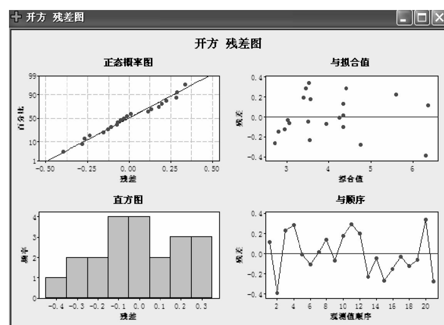


图4-110　变换后的残差分析图

在应用图4-109中的拟合方程时需要注意的是此时的Y值是氨损失量开方之后的值，在用于预测和优化时应该将结果平方之后使用。

4.2.4　咖啡好喝，不能多喝——非线性回归

变量之间除了存在线性关系之外，往往还存在平方或立方等非线性关系，非线性的模型往往可以更好地拟合实际数据。在Minitab中目前只能模拟两个变量之间非线性模型，即一个自变量和一个因变量。如果我们验证了一个自变量和因变量之间存在非线性关系，而因变量还和其他自变量存在线性关系，则可以将存在非线性关系的自变量按照非线性关系变换数据之后与 其他线性关系的自变量一起拟合线性关系。

某速溶咖啡品牌在制作过程中除了原料咖啡豆之外，还需要添加一种辅料来改善咖啡的口感，但同时又会增加咖啡因的含量。这种辅料需控制一定的添加量，添加到一定程度它会和咖啡豆中的另一种成分发生化学反应，从而导致咖啡因含量的降低。品牌商现想了解如何添加辅料可以既可以改善口感，又可以降低咖啡因含量。

经过一段时间的试验和数据收集，咖啡因含量和辅料的添加量数据如图4-111所示。

↓	C1	C2
	咖啡因含量	辅料
1	0.81740	3.01589
2	0.27346	3.58346
3	0.70578	3.11077
4	0.34237	3.81030
5	2.68945	4.36595
6	0.80550	3.03799
7	0.40824	3.49631
8	0.26911	3.33284
9	2.51762	4.39779
10	0.34187	3.61898
11	3.33540	4.50958
12	0.39860	3.16975
13	0.57638	3.85238
14	0.20345	3.66830
15	0.81752	3.95858
16	0.62691	3.08560
17	0.76066	3.07924

图4-111　咖啡因含量和辅料的添加量数据

从数据观察咖啡因的含量与辅料的添加量并不存在明显的线性关系，下面分析二者之间的关系。

（1）查看散点图

打开“咖啡因含量.MPJ”文件，将“咖啡因含量”作为Y 变量；“辅料”作为X 变量。在Minitab中制作散点图，设置如图4-112所示。



图4-112 设置散点图

单击【确定】按钮，生成的散点图如图4-113所示。

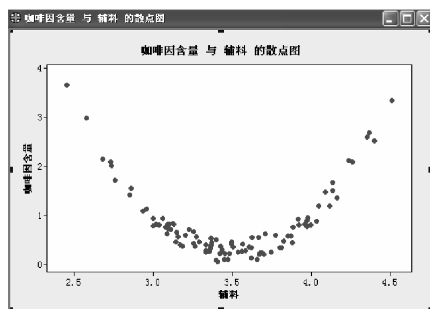


图4-113 咖啡因与辅料散点图

从图中可以看出咖啡因含量与辅料存在明显的曲线关系，下面对咖啡因含量和辅料之间进行拟合。

（2）非线性拟合

在Minitab中选择【统计】|【回归】|【拟合线图】选项，如图4-114所示。



图4-114　【拟合线图】选项

显示【拟合线图】对话框，在【响因】下拉列表框中选择【咖啡因含量】选项，在【预测变量】下拉列表框中选择【辅料】选项，在【回归模型】选项组中选择【二次】单选按钮，如图4-115所示。

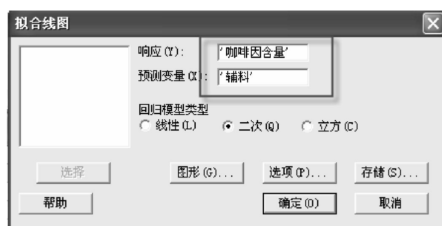


图4-115　设置二次拟合曲线

单击【图形】按钮，显示【拟合线图-图形】对话框。

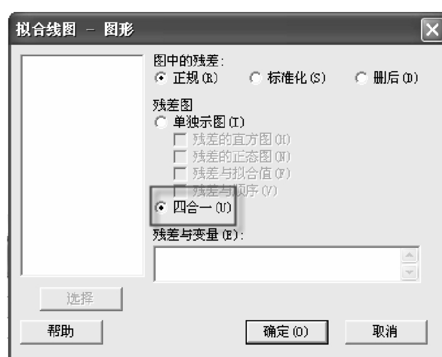
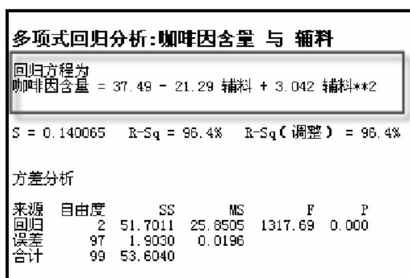


图4-116　【拟合线图-图形】对话框

单击【确定】按钮，分析结果如下4-117所示。



多项式回归分析：咖啡因含量 与 辅料

回归方程为
咖啡因含量 = 37.49 - 21.29 辅料 + 3.042 辅料**2

S = 0.140065 R-Sq = 96.4% R-Sq（调整） = 96.4%

方差分析

来源	自由度	SS	MS	F	P
回归	2	51.7011	25.8505	1317.69	0.000
误差	97	1.9030	0.0196		
合计	99	53.6040			

图4-117　分析结果

生成的拟合曲线图如图4-118所示。

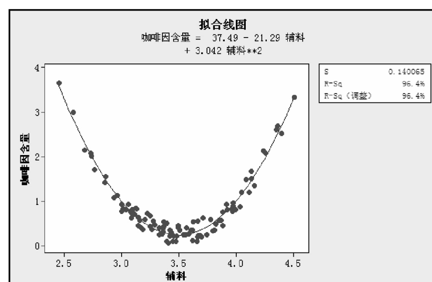


图4-118　生成的拟合曲线图

生成的残差四合一图如图4-119所示。

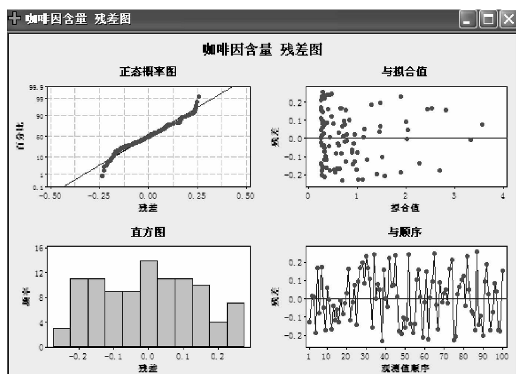


图4-119　生成的残差四合一图

从分析结果来看总体拟合良好，拟合方程为：

咖啡因含量＝“37.49-21.29辅料＋3.042辅料**2”，辅料的添加量大约在3.5左右时咖啡因含量最低。

残差分析四合一图均处于正常状态。

Miss　Ju：“我可以理解回归分析是在建立数学模型的吗？我们前面讨论LINGO、Crystal　Ball和Excel在进行规划求解及最优化求解时都需要模型，求解模型其实是应该在前面的工作，只是前面例子中的模型方程已提供给我们了，对吗？”

Mr Shu：“没错，其实按照数据分析的顺序我们应该求解出模型后进行优化或预测，不过也没有绝对的前后顺序。”

Miss　Ju：“我记得你前面提过回归分析除了计量型（连续性）数据外，离散型数据好像也可以进行回归，对吗？”

Mr Shu：“是的，前面列举的简单线性回归、多元线性回归和非线性回归都是计量型数据的　回归。在Minitab菜单中还有‘逐步’和‘最佳子集’等方法均是用来筛选自变量的一些方法，离散型数据的回归应用的比较少，我们简单了解一下Logistic回归。”

4.2.5　预防心血管疾病从减肥开始——二值Logistic回归分析

前面讨论的都是连续性变量的回归分析，离散型变量回归指对离

散型数据进行回归分析。离散型数据通常指“合格”、“不合格”、“增加”和“减少”等变量，这一类只取对立两值数据的回归问题称之为“离散变量的回归分析”。

在假设检验中离散型数据通常检验的是比例，离散数据的回归分析模型对于比例要使用Logistic变换，所以这种回归分析又称为“离散变量的Logistic回归”。该类回归分析又可以分为二值Logistic回归分析、名义值的Logistic回归分析和有序样本的Logistic回归分析，我们重点了解二值Logistic回归分析。

二值Logistic回归分析指因变量只有“是”或“否”两类。

肥胖是人类的健康杀手之一，通常使用体重指数来定义，即体重指数＝体重（公斤）／身高（米）^{2} 。例如，身高1.8米的人，体重为90公斤，则其体重指数为90/（1.8×1.8）＝27.8。按照通常的惯例，认为体重指数大于25以上属于肥胖。为了研究肥胖是否与患心血管疾病有关，某研究机构对某高校的学生进行了体检并记录了其中体重超重者是否患有心血管疾病。该数据如图4-120所示。

↓	C1	C2	C3	C4	C5
	体重指数	患病人数	未患病人数	总人数	患病率
1	25	68	42	110	0.618182
2	26	55	38	93	0.591398
3	27	66	20	86	0.767442
4	28	32	10	42	0.761905
5	29	21	7	28	0.750000
6	30	25	4	29	0.862069
7					

图4-120　体重与患病人数数据

从患病率数据来看，似乎体重指数越高，患病率越高。现在根据给定的体重指数是否能估算出患心血管病率的多少？我们需要对数据进行回归分析。

在Minitab中的操作步骤如下。

（1）二进制回归设置

打开“心血管疾病.MPJ”文件，在Minitab中选择【统计】|【回归】|【二进制Logistic回归】选项，如图4-121所示。



图4-121　【二进制Logistic回归】选项

显示【二进制Logistic回归】对话框，如图4-122所示。

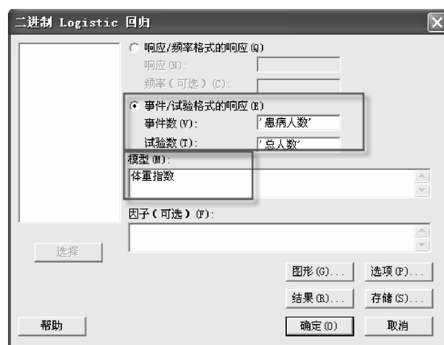


图4-122　【二进制Logistic回归】对话框

选择【事件/试验格式的响应】单选按钮，在【事件数】文本框中输入“患病人数”，【试验数】文本框中输入“总人数”。注意模型中输入的内容需与数据头列的名称保持一致；否则系统会报错，本例中输入“体重指数”。

在二进制Logistic回归分析时数据也可以如图4-123所示排列。这

是按照响应／频次格式对是否患病使用0和1记录，患有心血疾病的记录为1；否则为0。

在执行二进制Logistic回归分析时，需要按照响应／频次分析。可以选择图4-122中的第1种格式，即【响应／频率格式】选项。其排列格式如图4-123所示，且分析结果一致。

C6	C7	C8
体重系数	频数	是否患病
25	68	1
25	42	0
26	55	1
26	38	0
27	66	1
27	20	0
28	32	1
28	10	0
29	21	1
29	7	0
30	25	1
30	4	0

图4-123　二进制Logistic回归数据排列格式

（2）二进制回归分析结果

单击【确定】按钮，分析结果如图4-124所示。

二进制 Logistic 回归：患病人数，总人数 与 体重指数

连结函数:Logit

响应信息

变量	值	计数
患病人数	事件	267
	非事件	121
总人数	合计	388

Logistic 回归表

自变量	系数	系数标准误	Z	P	优势比	95% 置信区间 下限	95% 置信区间 上限
常量	-6.03235	2.08929	-2.89	0.004			
体重指数	0.256997	0.0789324	3.26	0.001	1.29	1.11	1.51

图4-124　二进制回归分析结果

（3）解读二进制回归分析结果

这里的分析结果与连续性数据分析结果不同，它未给出具体的拟合方程。但分析结果中有两个系数，分别为-6.03235和0.236997。常量和体重系数的检验概率*P* 值均小于0.05，说明模型是显著的。

分析结果给出的是Logistic函数的形式，即：

$$y=\ln(\frac{p}{1-p})=0.257\times\text{体重指数}-6.032$$

这表示体重指数增加一个单位，$\ln(\frac{p}{1-p})$ 将增加0.257，即$\frac{p}{1-p}$ 增加$e^{0.257}$ ＝1.293倍。

与连续性回归方程不同的是二进制Logistic回归分析需要进行Logistic函数转换。

分析结果中除了拟合方程外，还有拟合检验方法的*P* 值，如图4-125所示。

拟合优度检验

方法	卡方	自由度	P
Pearson	3.88627	4	0.422
偏差	3.86500	4	0.425
Hosmer-Lemeshow	3.37232	3	0.338

图4-125　拟合检验方法分析结果

在这里检验的零假设H0为拟合无显著差异，备择假设H1为拟合有显著差异。

3种方法的*P* 值均大于0.05，选择备择假设。即拟合有显著差异，拟合结果令人满意。

Minitab还会根据拟合方程计算预测频次数据，本例中的列1～列5分别代表体重指数为25～30的6组数据。由于体重指数为30的观测值中有小于6的数据，因此5组和6组合并为一组，即患病人数为21＋25＝46；未患病人数为4＋7＝11。预测频次数据，预测值如图4-126所示。

观测和期望频率表:
(有关 Pearson 卡方统计量，请参阅 Hosmer-Lemeshow 检验)

值	组 1	2	3	4	5	合计
事件						
观测值	68	55	66	32	46	267
期望值	65.7	61.1	61.3	32.0	47.0	
非事件						
观测值	42	38	20	10	11	121
期望值	44.3	31.9	24.7	10.0	10.0	
合计	110	93	86	42	57	388

图4-126　预测值

根据预测频次数可以计算出预测患病率，如体重指数为25时的预测患病率为65.7/(65.7＋44.3)＝0.597，其他数据依此类推。

在【二进制Logistic回归】对话框中单击【存储】按钮，显示【二进制Logistic回归-存储】对话框。选择【事件概率】复选框，如图4-127所示。

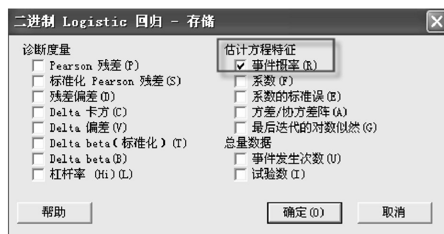


图4-127　【事件概率】复选框

单击【确定】按钮，事件概率值在Minitab表格中显示，如图4-128所示。

C7	C8	C9
频数	是否患病	事件概率1
68	1	0.596903
42	0	0.656914
55	1	0.712298
38	0	0.761981
66	1	0.805427
20	0	0.842582

图4-128　概率计算结果

计算结果与手动计算结果一致，从预测的数据来看我们是否可以由此得出这样的结论，即体重指标越大者发生患心血管疾病的概率越大。除了在二进制回归分析的结果还有相关性度量分析结果，如图4-129所示。

相联度量:
（响应变量与预测概率之间）

配对	数字	百分比	度量结果综述	
一致	16025	49.6	Somer 的 D	0.20
不一致	9449	29.2	Goodman-Kruskal Gamma	0.26
结	6833	21.2	Kendall 的 Tau-a	0.09
合计	32307	100.0		

图4-129　相关性度量分析结果

我们本次抽样的样本量为267个“患心血管疾病”患者和121个“未患心血管疾病”，如果两两配对，则可以形成121×267＝32　307个。其中符合体重指标越大，患心血管疾病的概率高这一规律的有16　025对，占总配对数的49.6%；不符合有9　449对，占总配对数的29.2%；其他为6 833，占总配对数的21.2%。

【Somer的D】、【Goodman-Kruskal　Gamma】和【Kendall的Tau-a】都是一致性结果的度量，三者的分子都是一致配对数减去不一致配对数，本例中为16 025-9 449＝6 576。分母则不相同，【Somer的D】分母为总配对数，即32 307，一致性为6 576/32 307＝0.20；【Goodman-Kruskal　Gamma】分母为扣除“结”的总配对数，即32 307-6 833＝25 474，一致性为6 576/25 474＝0.26；【Kendall的

Tau-a】分母为所有可能的观测值的配对，(267＋121)×(267＋121-1)/2＝75078，即一致性为6 576/75 078＝0.09。

计算出来的3个一致性统计值分别为0.20、0.26和0.09，数值相对比较低。3个数据值为0～1，值越大，表示模型的预测能力越强。

这些数据还不足以证明体重指数大患心血管疾病的概率就高的结论，这与我们抽取的样本 量大小有关系。抽样的样本量越大，越能准确地反映实际情况。

Miss Ju："我可以理解二进制的回归分析相当于多样本的比例检验吗？因为我们前面讨论过单样本和双样本的比例检验，而二进制的回归分析的样本数要比它们都多且检验的也是比例。"

Mr Shu："可以这样关联学习，但是不完全等同。在我们很多数据分析中都应用假设检验。在处理二进制的回归分析时也把计数型数据，如满意和不满意等转换成比例这类计数型数据"。

离散型变量的回归分析还有名义值的Logistic回归分析和有序样本的Logistic回归分析，作为数据分析的入门学习不再深入。在4.1节中我们了解的大多是单个X 与单个Y 之间的分析，结合前面案例我们将回归分析做一个简短的小结，如表4-9所示。

表**4-9**　回归分析方法汇总

数据类型	单个Y	单个Y	单个Y	单个Y	单个Y
	单个X	多个X	多个X	多个X	多个X
	X计量	X计量	X计数	X计量	X计数
	Y计量	Y计量	Y计量	Y计数	Y计数
分析方法	相关/简单回归	多元回归	双因子或多因子方差分析	多元逻辑回归	多元逻辑回归

前面所有讨论的分析方法都是针对单个Y 的分析，如果数据中出

现多个Y ，则要使用多元分类数据分析。这类分析相对复杂，下面我们重点介绍与多元分类数据分析相关且有趣的聚类分析。

4.3 人以类聚，物以群分——聚类分析

Mr Shu：“物以类聚、人以群分说的就是聚类，接下来我们要讨论聚类分析。”

Miss Ju：“但是这个类应该怎么分呢？像我们经常说的发达国家和发展中国家等就是把国家按照经济发展的程度划分类别的，对吗？”

Mr Shu：“是的，国家的分类问题在冷战时期可以按国家政体形式和经济发展水平分为第一世界、第二世界和第三世界国家。在当今世界，国际形势早已不是几十年前的状态。国与国之间关系错综复杂，将国家分类需要考虑的因素非常多，如国家政体模式、经济发展水平与模式、军事力量对比，以及民族关系等。国家分类必须将这些重要因素综合考虑，可以从每一种因素 的数据入手充分利用数据，客观描述国与国之间的关系。将诸多特征相似的国家分在一组，不相似的国家分在另外的组中，这也是一个典型的聚类分析的问题。

Miss Ju：“就像在数据分析界非常有名的案例啤酒与尿布也是发现了购买这两类产品的同一属性，即年轻的父亲，也可以算是聚类分析吗？”

Mr Shu：“没错，其实在超市中啤酒与尿布的现象比比皆是。为什么这个案例只发生在沃尔玛？这与其背后强大的计算机及数

据分析团队密不可分，现在很多超市也是按照沃尔玛研究的购物篮分析来优化其商品布置。”

Miss Ju：“看来很多的数据分析方法早就有了应用。”

Mr Shu：“当然，不过数据分析只是改善管理的手段和工具，真正要改善管理还在于管理者的头脑和思路。现在暂不关心管理类的方法，我们关心的是如何进行聚类分析。由于聚类分析涉及复杂的矩阵计算且每种聚类的方法都有复杂的方程式，因此作为数据分析的入门级学习了解这个小节的内容即可。”

聚类分析又称“群分析”，是分类学的一种基本方法。所谓“类”，通俗地讲即由具有相似性元素构成的集合。聚类分析也是多元统计学中应用极为广泛的一种重要方法。

分类问题在经济、医学及科学研究中十分常见，如超市商品的种类繁多。需要根据商品的用途、价格档次及产地等多方面的变量因素分成不同的组别，仅仅利用单变量因素分布不足以全面且综合地描述商品的类别。应多种因素同时考虑并对不同品种之间的关系给出定量化的描述，然后制定分组规则，按照实际的需要分类可以利用聚类分析的方法。

聚类分析是对研究样品或指标进行分类的一种多元统计方法，这种方法正处于发展阶段，理论上仍很不完善。但是由于它能够解决许多实际问题，因此在很多具体问题及应用建模中得到了重视，特别是和其他统计方法结合起来使用效果更好。

聚类分析仍是一种解释数据的方法，要得到一个客观且综合的聚类分析结果必须经过多次不同方法的实验；同时辅助其他方法，例如统计图形和机理分析等。聚类分析过程不仅是一门分类的科学方法，更像是一门分类的艺术。

在Minitab中提供了如下3种聚类分析方法。

（1）观测值聚类：使用分层聚类法，以所有观测值都是独立且每个都形成自己的聚类为出发点。该方法类适用于小表，最多只能有几千行，并按树结构的层次顺序合并行。在Minitab中的树状图是一种动态响应图，在建立之后可以选择所需的聚类数。聚类数不同，树状图也不 相同。

（2）变量聚类：分层聚类法，变量聚类的作用之一是减少变量的维度。

（3）K均值聚类：此过程根据MacQueen算法。K均值聚类适用于较大表，多达几十万行。首先K均值聚类将对聚类种子点进行一个非常完善的预测，然后开始迭代。交替执行两个操作，即将点指定给聚类和重新计算聚类中心。用户必须指定聚类数，然后才能开始这一过程。

4.3.1 美好一天从早餐开始——观测值聚类分析

早餐是补充人体能量的最要环节，一般情况下人们都会选择营养丰富又少脂肪的粗粮作为早餐。可供早餐选择的种类比较多，人们应该如何选择适合自己的食物？

营养研究学人员检测了11种常食用的早餐的营养成本并记录，数据如图4-130所示。

↓	C1-T	C2	C3	C4	C5	C6
	品牌	蛋白质	碳水化合物	脂肪	卡路里	维生素A
1	核桃	6	19	1	110	0
2	葡萄坚果	3	23	0	100	25
3	超级酥糖	2	26	0	110	25
4	开心果	6	21	0	110	25
5	爆米香	2	25	0	110	25
6	葡萄干	3	28	1	120	25
7	小麦片	2	24	0	110	100
8	小麦粉	3	23	1	110	25
9	芝麻	3	23	1	110	100
10	爆米花	1	13	0	50	0
11	糖爆玉米花	1	26	0	110	25
12	风味糖	2	25	0	110	25

图4-130　食物的营养成本数据

观测值聚类分析的操作步骤如下。

（1）选择观测值聚类分析

打开“早餐.MPJ”文件，在Minitab中选择【统计】|【多变量】|【观测值聚类】选项，如图4-131所示。

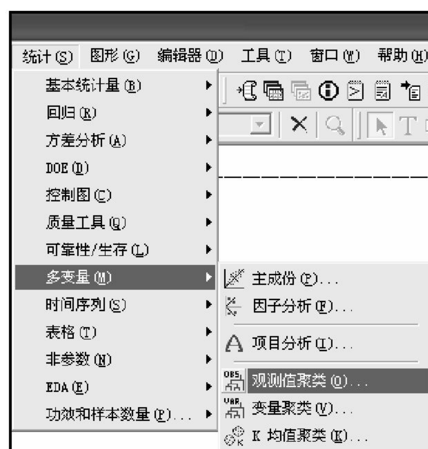


图4-131　【观测值聚类】选项

（2）设置观测值聚类分析

显示【观测值聚类】对话框，如图4-132所示。



图4-132　【观测值聚类】对话框

在【变量或距离矩阵】下拉列表框中选择要聚类的变量，本例要按照营养成分进行聚类。因此选择C2～C6列，即从蛋白质到维生素A。

前面提到观测值聚类按照分层距离大小分类，测量距离有一定法则。【联结法】下拉列表框用于指定距离法则，在Minitab中提供了平均、质心、最长距离、简单平均、中间距离、最短距离及离差平方和7种法则。每种法则聚类之后的结果不同，本例中选择【最长距离法】法则。

有了距离的法则，还需要计算距离的具体公式。Minitab中提供了5种计算方法，即欧氏距离Euclidean、平方欧氏距离Squared Euclidean、Pearson距离、Pearson平方和Manhattan，每种计算距离的公式也不一样。本例中在【距离量度】下拉列表框中选择【平方欧氏距离Squared Euclidean】选项。

在计算距离之前将所有变量转换为公共尺度，方法为减去平均值并除以标准差，与概率分布中计算Z值的方法一样。如果变量使用不同的单位且计划最大限度地降低尺度差异带来的影响，则是一种很好的做法，本例选择【标准化变量】复选框。

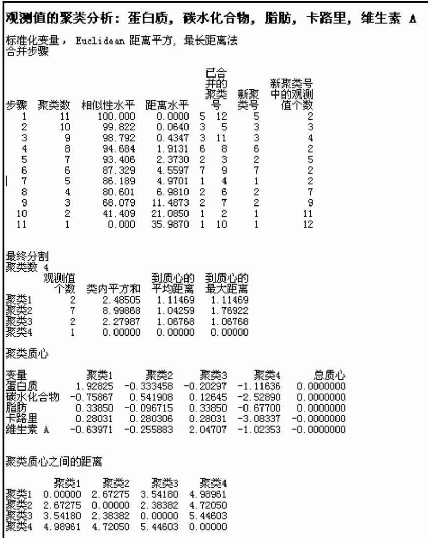
选择【聚类数】单选按钮，并指定最终分出来的类别数为4。如

果选择【相似性水平】单选按钮，则按照相似性水平分类。

选择【显示树状图】复选框。

（3）解读观测值聚类分析结果

单击【确定】按钮，分析结果如图4-133所示。



观测值的聚类分析：蛋白质，碳水化合物，脂肪，卡路里，维生素 A

标准化变量， Euclidean 距离平方，最长距离法
合并步骤

步骤	聚类数	相似性水平	距离水平	已合并的聚类号		新聚类号	新聚类号中的观测值个数
1	11	100.000	0.0000	5	12	5	2
2	10	99.822	0.0640	3	5	3	3
3	9	98.792	0.4347	3	11	3	4
4	8	94.684	1.9131	6	8	6	2
5	7	93.406	2.3730	2	3	2	5
6	6	87.329	4.5597	7	9	7	2
7	5	86.189	4.9701	1	4	1	2
8	4	80.601	6.9810	2	6	2	7
9	3	68.079	11.4873	2	7	2	9
10	2	41.409	21.0850	1	2	1	11
11	1	0.000	35.9870	1	10	1	12

最终分割
聚类数 4

	观测值个数	类内平方和	到质心的平均距离	到质心的最大距离
聚类1	2	2.48505	1.11469	1.11469
聚类2	7	8.99868	1.04259	1.76922
聚类3	2	2.27987	1.06768	1.06768
聚类4	1	0.00000	0.00000	0.00000

聚类质心

变量	聚类1	聚类2	聚类3	聚类4	总质心
蛋白质	1.92825	-0.333458	-0.20297	-1.11636	0.0000000
碳水化合物	-0.75867	0.541908	0.12645	-2.52890	0.0000000
脂肪	0.33850	-0.096715	0.33850	-0.67700	0.0000000
卡路里	0.28031	0.280306	0.28031	-3.08337	-0.0000000
维生素 A	-0.63971	-0.255883	2.04707	-1.02353	-0.0000000

聚类质心之间的距离

	聚类1	聚类2	聚类3	聚类4
聚类1	0.00000	2.67275	3.54180	4.98961
聚类2	2.67275	0.00000	2.38382	4.72050
聚类3	3.54180	2.38382	0.00000	5.44603
聚类4	4.98961	4.72050	5.44603	0.00000

图4-133　聚类分析结果

树状图如图4-134所示。

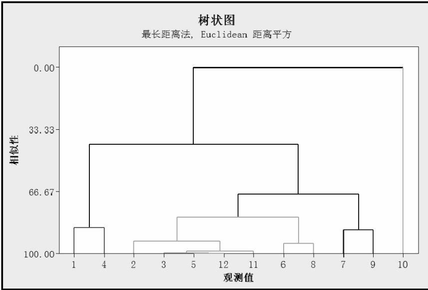


图4-134　聚类分析树状图

树状图以树形图的形式显示合并表中的信息，在我们的示例中早餐食品1和4组成第1个聚类；早餐食品2、3、5、12、11、6和8组成第2个聚类；早餐食品7和9组成第3个聚类：早餐食品10组成第4个聚类。

4.3.2　海拔是否影响血压——变量聚类分析

西藏在中国的高原地带，空气新鲜且风景优美，是许多海拔较低的大城市的旅游爱好者向往的圣地。现有研究者想确定高海拔对人体的血压是否有影响，他们对39名年龄在21岁以上的成人进行了跟踪研究，这39人从高海拔的西藏地区迁移到平原地区生活。研究者记录了他们的年龄；迁移年数；体重（公斤）；身高（毫米）；下巴、前臂和小腿的皮褶厚度（毫米）；脉搏，以及心脏的收缩和舒张，收集的数据如图4-135所示。

↓	C1 年龄	C2 年数	C3 体重	C4 高度	C5 下巴	C6 前臂	C7 小腿	C8 脉搏
1	21	1	71.0	1629	8.0	7.0	12.7	88
2	22	6	56.5	1569	3.3	5.0	8.0	64
3	24	5	56.0	1561	3.3	1.3	4.3	68
4	24	1	61.0	1619	3.7	3.0	4.3	52
5	25	1	65.0	1566	9.0	12.7	20.7	72
6	27	19	62.0	1639	3.0	3.3	5.7	72
7	28	5	53.0	1494	7.3	4.7	8.0	64
8	28	25	53.0	1568	3.7	4.3	0.0	80
9	31	6	65.0	1540	10.3	9.0	10.0	76
10	32	13	57.0	1530	5.7	4.0	6.0	60
11	33	13	66.5	1622	6.0	5.7	8.3	68
12	33	10	59.1	1486	6.7	5.3	10.3	72
13	34	15	64.0	1578	3.3	5.3	7.0	88
14	35	18	69.5	1645	9.3	5.0	7.0	60
15	35	2	64.0	1648	3.0	3.7	6.7	60
16	36	12	56.5	1521	3.3	5.0	11.7	72
17	36	15	57.0	1547	3.0	3.0	6.0	84
18	37	16	55.0	1505	4.3	5.0	7.0	64
19	37	17	57.0	1473	6.0	5.3	11.7	72

图4-135　收集的数据

（1）选择变量聚类分析

打开“西藏.MPJ”文件，在Minitab中选择【统计】|【多变量】|【变量聚类】选项，如图4-136所示。

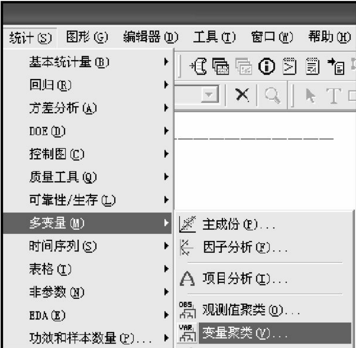


图4-136 【变量聚类】选项

（2）设置变量聚类分析

显示【变量聚类】对话框，如图4-137所示。

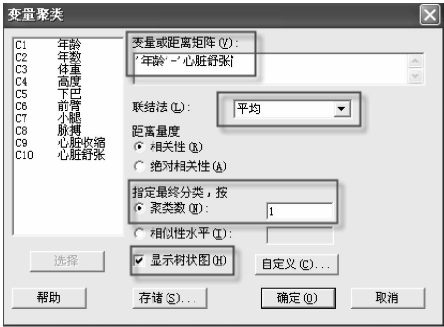


图4-137 【变量聚类】对话框

在【变量或距离矩阵】下拉列表框中选择C1～C10单元格，在【联结法】下拉列表框中选择【平均】选项。聚类数设置为1，选择【显示树状图】复选框。

（3）解读变量聚类分析结果

单击【确定】按钮，分析结果如图4-138所示。

结果：西藏.MTW

变量的聚类分析：年龄，年数，体重，高度，下巴，前臂，小腿，脉搏，...

相关系数距离，类平均法
合并步骤

步骤	聚类数	相似性水平	距离水平	已合并的聚类号		新聚类号	新聚类号中的观测值个数
1	9	86.7763	0.264474	6	7	6	2
2	8	79.4106	0.411787	1	2	1	2
3	7	78.8470	0.423059	5	6	5	3
4	6	76.0682	0.478636	3	9	3	2
5	5	71.7422	0.565156	3	10	3	3
6	4	65.5459	0.689082	3	5	3	6
7	3	61.3391	0.773218	3	8	3	7
8	2	56.5958	0.868085	1	3	1	9
9	1	55.4390	0.891221	1	4	1	10

图4-138　变量聚类分析结果

树状图如图4-139所示。

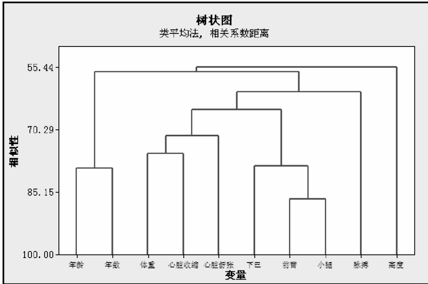


图4-139　聚类分析树状图

在此示例中下巴、前臂和小腿皮褶厚度的测量值相似，因此合并它们；年龄和迁移年数这两个变量相似，如果对象在某个特定年龄迁移，则这些变量可能会包含相似信息并且可以合并；重量和两个血压测量值相似，系统会自动保留一个变量。即将重量作为独立变量保留，将血压测量值合并为一个。

4.3.3　为熊猫分类——K均值聚类分析

熊猫是中国的国宝，熊类科学家需要记录其成长并分析成长特性以便更好地保护它们。目前大部分熊猫圈养在动物园内，研究人员需要根据其身体状况分类决定哪些熊猫已足够强壮到可以适应野外生活。

研究人员需要测量熊猫的总长、头长、总重、头重、颈围和胸围，共测量了143只熊猫并将其分为小中大3类，其中，第2只、第78只和第15只分别属于这3类大小的典型代表。

收集的熊猫数据如图4-140所示。

C1	C2	C3	C4	C5	C6	C7	C8
ID	年龄	月份	性别	头长	头宽	颈围	长度
39	19	7	1	10.0	5.0	15.0	45.0
41	19	7	2	11.0	6.5	20.0	47.5
41	20	8	2	12.0	6.0	17.0	57.0
41	23	11	2	12.5	5.0	20.5	59.5
41	29	5	2	12.0	6.0	18.0	62.0
43	19	7	1	11.0	5.5	16.0	53.0
43	20	8	1	12.0	5.5	17.0	56.0
45	55	7	1	16.5	9.0	28.0	67.5
45	67	7	1	16.5	9.0	27.0	78.0
48	81	9	1	15.5	8.0	31.0	72.0
69	*	10	1	16.0	8.0	32.0	77.0
83	115	7	1	17.0	10.0	31.5	72.0
83	117	9	1	15.5	7.5	32.0	75.0
83	124	4	1	17.5	8.0	32.0	75.0
83	140	8	1	15.0	9.0	33.0	75.0
91	104	8	2	15.5	6.5	22.0	62.0

图4-140　收集的熊猫数据

（1）设置初始分隔列

与变量聚类和观测值聚类不同，K均值聚类分析需要指定初始分隔列。将3只种子熊猫指定为1＝小、2＝中和3＝大，并将其余熊猫指定为0，以指示初始聚类成员。

打开“熊猫.MPJ”文件，在Minitab中选择【计算】|【产生模板化数据】|【简单数集】选项，显示【简单数集】对话框，如图4-141所示。

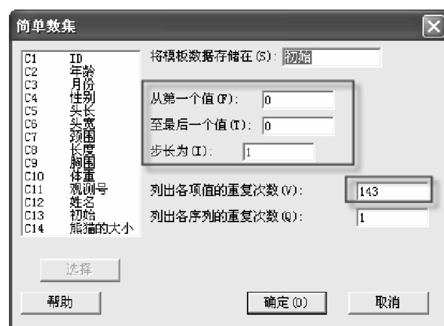


图4-141 【简单数集】对话框

在【将模板数据存储在】文本框中输入“初始”，将产生的数据放在初始列。将初始列中的第2行、第78行和第15行分别设置为1、2和3，如图4-142所示。

C11 观测号	C12 初始
1	0
1	1
2	0
3	0

图4-142 设置初始列

单击【确定】按钮。

（2）选择K均值聚类分析

在Minitab中选择【统计】|【多变量】|【K均值聚类】选项，如图4-143所示。

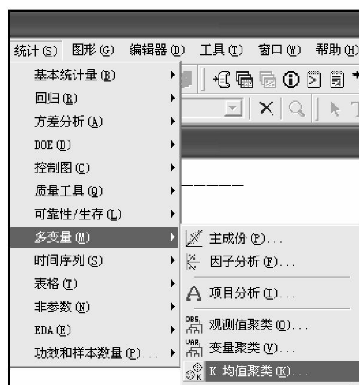


图4-143 【K均值聚类】选项

（3）设置K均值聚类分析

显示【K均值聚类】对话框，如图4-144所示。



图4-144 【K均值聚类】对话框

在【变量】下拉列表框中输入C5～C10，按照初始分隔列分割。选择【初始分割列】单选按钮和选择【标准化变量】复选框。单击【存储】按钮，显示【K均值聚类-存储】对话框，如图4-145所示。

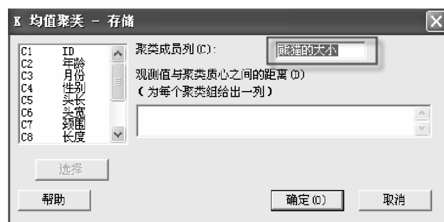


图4-145 【K均值聚类-存储】对话框

在【聚类成员列】文本框中输入“熊猫的大小”，单击【确定】按钮，分析结果如图4-146所示。

K 均值聚类分析：头长，头宽，颈围，长度，胸围，体重

标准化变量

最终分割

聚类数 3

	观测值个数	类内平方和	到质心的平均距离	到质心的最大距离
聚类1	41	63.075	1.125	2.488
聚类2	67	78.947	0.997	2.048
聚类3	35	65.149	1.311	2.449

聚类质心

变量	聚类1	聚类2	聚类3	总质心
头长	-1.0673	0.0126	1.2261	-0.0000
头宽	-0.9943	-0.0155	1.1943	0.0000
颈围	-1.0244	-0.1293	1.4476	-0.0000
长度	-1.1399	0.0614	1.2177	0.0000
胸围	-1.0570	-0.0810	1.3932	-0.0000
体重	-0.9460	-0.2033	1.4974	-0.0000

聚类质心之间的距离

	聚类1	聚类2	聚类3
聚类1	0.0000	2.4233	5.8045
聚类2	2.4233	0.0000	3.4388
聚类3	5.8045	3.4388	0.0000

图4-146 K均值聚类分析结果

分析结果存储在Minitab中的“熊猫的大小”列中，如图4-147所示。

C11 观测号	C12 初始	C13 熊猫的大小
1	0	1
1	1	1
2	0	1
3	0	2
4	0	2
1	0	1
2	0	1
1	0	3
2	0	3
1	0	3
1	0	3
1	0	3
2	0	3

图4-147 “熊猫的大小”列

在Minitab表格中出现了新的一列，即“熊猫的大小”并给出了熊猫分类的大小，K均值聚类将143只熊猫划分为41只小熊猫、67只中熊猫和45只大熊猫。

聚类分析的基本思想是对于位置类别的样本或变量，依据相应的定义将其分为若干类。分类过程是一个逐步减少类别的过程，在每一个聚类层次必须满足“类内差异小，类间差异大”原则，直至全部归为一类。聚类的方法有很多，每种软件的操作和分类也不一样。如Minitab按照“观测值聚类”、“变量聚类”和“K均值聚类”等进行，而JMP则按照“层次聚类”、“K均值聚类”和“正态混合聚类”进行。

不管使用何种软件或工具，聚类的目的都是为了根据已知数据计算各观察个体或变量之间亲疏关系的统计量（距离或相关系数）。距离的计算方法又有多种，选择不同的计算方法聚类的结果也不同。

第5章 要里子，也要面子 数据展现的艺术

5.1 哪种水果更好卖

5.2 书店利润最大化

5.3 非诚勿扰——最佳男友模型

Mr Shu：“入门级的数据分析我们讨论得差不多了，你有什么想法？”

Miss Ju：“这还叫入门级啊，够我消化好一阵。”

Mr Shu：“真正高大上的数据分析需要有IT知识构架，并且会编写程序，即要有‘码农’的职业精神。IT加统计加业务才是真正的高手，我们这样只能算是入门了。”

Miss Ju：“哇喔，我最烦的就是写程序，上学时一个C语言就够我折腾了。”

Mr Shu：“我们有了前面的数据分析、数据模型和预测等这

些‘硬货’，还需要一些上台面的‘软货’。”

Miss Ju：“什么叫软货？”

Mr Shu：“有了数据分析的结果，我们不可能直接拿着Minitab或Excel汇报成果？领导不关心技术细节，他们需要的一般都是结果。软货就是数据的展示，即我们下面要介绍的Xcelsius软件，俗称‘水晶易表’。”

Miss Ju：“水晶易表？与水晶球有关系吗？”

Mr Shu：“没有关系，是中文翻译的凑巧。你听说过SAP公司的BI商业智能吧？它是现在比较火热的概念。Xcelsius是BI的重要成员。并且完全就是傻瓜式数据可视化，操作非常简单。”

Miss Ju：“简单？你别说了，前面你都说简单。”

Mr Shu：“这个是真的简单，只要你稍懂Excel就可以玩转它，用它瞬间可以制作出惊人的报告。”

Xcelsius是全球领先的商务智能软件商Business Objects的产品，Xcelsius是Excel与Celsius（摄氏度）两个单词的结合，表示Xcelsius比Excel更灵动并且像温度计一样是易于读取和易于使用的工具。

Xcelsius开始出自一名美国加州少年的天才与灵感，并从Flash制作、电子游戏开发、多媒体技术中吸取知识。他与父亲一起创业并合作完善Xcelsius产品，随后Xcelsius在美国IT界多次获得高级奖项。2005年Business Objects收购了Xcelsius的创始者Information公司，并作为商务智能产品中的重要组成部分。

Xcelsius于2006年推广至中国，随后Business Objects被SAP公司并购。并且将其命名为“水晶易表”，表明它是一种非常容易掌握的软

件。

商务智能主要有几个方面的应用，即数据查询、在线分析、报表工具及深层次分析。最具有代表性的是仪表盘的应用，利用仪表盘来动态表现数据，以帮助企业获得更多的商业洞察力。

我们介绍的重点是Xcelsius在报表工具中的应用，与传统的Excel图表展现不同的是Excel是静态的展现。需要动态展现一般都需要借助VBA，这就加大了难度，而Xcelsius只需要通过简单的拖动操作就可以实现用户与数据的互动。

Xcelsius在Excel表格模型基础上实现这些强大功能，有些功能可以在Xcelsius或Excel中实现。将二者的优势结合能做出非常精美的动态PPT文档，让我们的数据分析结果得到充分的展示。

Miss Ju：“你前面说只要有点Excel基础就可以很快完成惊人的报告，能不能先演示一个例子？”

Mr Shu：“一般来讲5分钟左右就可以完成一个简单的动态PPT。”

Miss Ju：“真的？我们演示一个示例。”

Mr Shu：“好的，通过Xcelsius制作动态仪表盘一般只需要3个步骤就可以完成与Excel数据的交互和可视化分析。即第1步导入已有的Excel电子表格；第2步通过交互式操作界面创建可视化分析，不需要额外的编程；第3步将分析结果直接嵌入到PowerPoint、PDF文件中或直接发送邮件。”

在安装成功Xcelsius后，单击【程序】|【Xcelsius】选项，显示的

Xcelsius界面如图5-1所示。

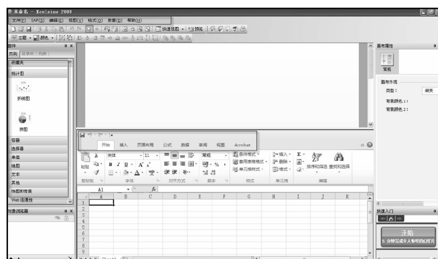


图5-1　Xcelsius界面

其中白色部分称为“画布工作区”即在这里实现编辑设计和可视化分析。这里类似画画，所以称之为“画布”。

白布的下方是Excel文件，既可以在Xcelsius中直接编辑Excel文件，也可以在外部将Excel文件编辑好之后导入至Xcelsius中。

工具区的上方是工具栏；左侧上半部分是【部件】面板，即我们需要关注的重点，其中有很多已经设计好的部件，分为【统计图】、【容器】、【选择器】、【单值】和【地图】等部件大类，在这些大类中还有子类，它们为数据的展示提供了多种选择。选中直接拖动需要的组件到画布就可以进行编辑，组件的大小可以根据需要拉伸。

【部件】面板下方是【对象浏览器】窗格，其中显示画布中的所有的部件。当部件较多且叠放时，往往不易选中。此时通过【对象浏览器】窗格选择，该窗格还可以隐藏和锁定选中的部件。

5.1　哪种水果更好卖

某家进口水果商想了解其进口的苹果、葡萄和香蕉近一年的销售

情况及趋势，以确定第2年的购买数量，为此他记录了每月水果的销售情况。在Xcelsius创建动态仪表盘的操作步骤如下。

（1）创建Excel表格

Excel数据模型是制作Xcelsius动态仪表盘的基础，一套精美的动态仪表盘需要设计合理的Excel作为支撑。进口水果商记录的原始销售数据如图5-2所示。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	名称	1	2	3	4	5	6	7	8	9	10	11	12
2	进口苹果	581	581	535	726	1133	1184	754	906	1059	1661	1030	800
3	进口葡萄	1346	745	1275	774	875	1673	806	2095	483	694	1745	927
4	进口香蕉	533	597	2359	912	980	1023	875	464	442	484	981	1090

图5-2　原始销售数据

由于我们在后期制作动态仪表盘时需要引用一些单元格区域，因此修正后Excel表，如图5-3所示。

将原表格下移至下部区域，复制表格中的前两行至表的前两行，在A4和A5单元格中分别输入“折线图”和“柱状图”。这个区域提供一个选择开关选择，用于折线图和柱状图之间的切　换，并在A4和A5单元格填充颜色。在Excel模型中填充颜色是为了划分数据区域，方便数据的选择和插入。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	名称	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
2	进口苹果	581	581	535	726	1133	1184	754	906	1059	1661	1030	800
3													
4	折线图												
5	柱状图												
6													
7	平均值	912	实际值										
8	最大值	1661	目标值										
9	最小值	535											
10													
11													
12													
13	名称	1	2	3	4	5	6	7	8	9	10	11	12
14	进口苹果	581	581	535	726	1133	1184	754	906	1059	1661	1030	800
15	进口葡萄	1346	745	1275	774	875	1673	806	2095	483	694	1745	927
16	进口香蕉	533	597	2359	912	980	1023	875	464	442	484	981	1090

图5-3　修正后的Excel模型

在A7～A9单元格中分别输入“平均值”、“最大值”和“最小值”，并在A7中输入公式“＝AVERAGE(B2:M2)”，在A8中输入公式“＝Max(B2:M2)”，在A9中输入公式“＝Min(B2:M2)”，在C7和C8单

元格中分别输入“实际值”和“目标值”用于呈现数据。

Excel模型的设计比较简单，设计完成后就可以导入Xcelsius。

（2）导入数据

启动Xcelsius软件，在菜单中选择【数据】|【导入】选项，如图5-4所示。



图5-4 【导入】选项

显示【导入】对话框，选择“水果销售数据.xls”文件，单击【确定】按钮。

（3）设置画布

由于需要动态查看3种进口水果的不同销售情况，因此在Xcelsius中动态选择的按钮比较多，主要是在【选择器】大类中我们可以根据需要选择。选择【标签式菜单】选项，并将其拖至画布中。拖动过程中鼠标会变成小的加号“＋”，添加标签后的画布如图5-5所示。

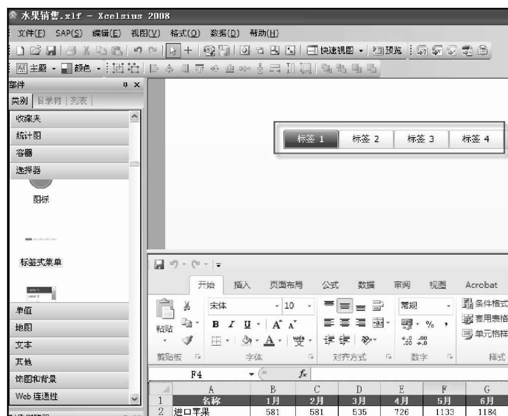


图5-5　添加标签后的画布

双击白色画布中的【标签式菜单】按钮，显示【标签式菜单1】对话框，如图5-6所示。



图5-6　【标签式菜单1】对话框

在【标题】下拉列表框中输入“进口水果”，也可以通过Excel单元

格进行引用。凡显示小图标处说明都可以进行Excel单元格引用。在【标签】下拉列表框中选择A14:A16单元格区域，即3种水果的名称。

在【数据插入】选项组的【插入类型】下拉列表框中包括【位置】、【标签】、【值】、【行】、【列】、【过滤后的行】和【状态列表】等7个选项，每种选项均会出现不同的插入方式，本例选择【行】选项。源数据为Excel表格中的数据，选择B14:M16单元格区域。目标即插入数据后存放的位置，选择B2:M2单元格区域。由于我们插入数据时的方向都是行，所以在【方向】选项组中选择【行】单选按钮。

上面仅仅是设置常规选项，设置完成后画布中的标签式菜单如图5-7所示。

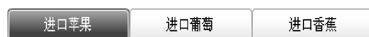


图5-7　设置常规选项后的标签式菜单

下面需要添加图形并设置，本例选择【折线图】选项，添加后的结果如图5-8所示。



图5-8　折线图

双击【折线图】，显示【折线图1】对话框，如图5-9所示。



图5-9　【折线图1】对话框

在【常规】选项卡【标题】选项组中的【统计图】下拉列表框中引用A2单元格内容，【副标题】、【类别轴】和【值轴】等下拉列表框均设置为空。

在【数据】选项组选择【按范围】单选按钮，并对引用需要显示的数据区域B1:M2。设置好后系统会自动在下方的【按系列】选项组中链接相应的单元格，并自动生成系列名称，将系列名称修正为“销售额”。

设置后的图形如图5-10所示。



图5-10　设置后的图形

菜单与折线图此时已可以实现联动，单击工具栏中的【浏览】按钮进入浏览状态，此时任意选择3种进口水果所对应的标签就可以动态查看其对应的折线图。

通过这个简单的案例我们讨论了Xcelsius部件中的两大类，即【选择器】和【统计图】。每种大类都有多个子类可供选择，如【选择器】可以选择【复选框】和【组合框】等选项；【统计图】可以选择【柱状图】和【饼图】等图形，其操作过程均大同小异。

Miss Ju：“这么说动态图表已设置完成了。”

Mr Shu：“是的，简单的动态图表已经设置完成了，可以保存成PPT或发送邮件了。”

Miss　Ju：“我们之前在Excel模型中设置的最大值和平均值还有折线与柱状之间又如何选择？”

Mr Shu：“下面我们就来讨论它稍深入的应用。”

（4）添加柱状图

拖动【部件】面板中的【柱状图】至画布中，设置与折线图一样，设置后的折线图与柱状图如图5-11所示。

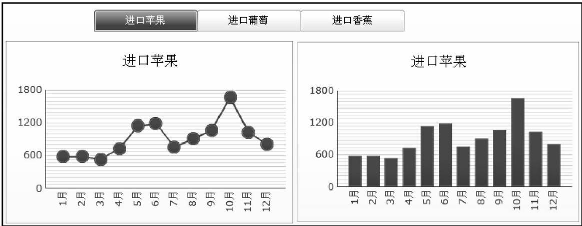


图5-11　设置后的折线图与柱状图

单击【浏览】按钮，折线图与柱状图可以同步动态变化。但我们不需要两个图形同时出现，所以需要对两个图形进行位置重叠。

按住Ctrl键在【对象浏览器】窗格中选择折线图和统计图，在菜单中选择【格式】|【相同大小】|【两者】选项。依次选择【格式】|【对齐】|【左对齐】选项及【格式】|【对齐】|【顶对齐】选项，如图5-12所示。



图5-12　重叠两个图形位置的选项

（5）添加选择器

我们还需要添加另一个选择来切换重叠后的图形。

在【选择器】大类中选择【单选按钮】选项，并将其拖动至画布中，如图5-13所示。

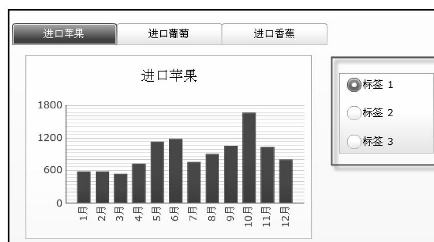


图5-13　添加单选按钮

设置单选按钮的属性，双击单选按钮显示【单选按钮1】对话框，如图5-14所示。



图5-14　【单选按钮1】对话框

在【标签】下拉列表框中选择A4:A5单元格区域，即Excel表格中存放折线图和柱状图文字的单元格，也可以单击 按钮后手动输入。在【插入类型】下拉列表框中选择【位置】选项，在【目标】下拉列表框中选择一个空单元格B5，在【方向】选项组中选择【水平】单选按钮。B5单元格用于存放该选择器赋予的值，选择折线图时被赋予数字“1”；选择“柱状图”时被赋予数字“2”。正是因为有了Excel单元格的数值才可以实现图形的切换。

设置完成后还需要将图形与Excel单元格建立联系。

在【对象浏览器】窗格中选择【折线图】选项，双击折线图，显示【折线图1】对话框，如图5-15所示。



图5-15　【折线图1】对话框

在【行为】选项卡中设置动态可见性中状态值与B5单元格连接，并设置代码为1，即B5单元格为数值1时折线图可见；同样将柱状图的动态可见性设置为代码2。单击【浏览】按钮可以非常方便地切换。

Miss Ju："太神奇了吧！这么简单的操作就实现了图表的切换？"

Mr Shu："是的，就这么简单！"

Miss Ju："剩下的工作是不是就是美化图表了？"

Mr Shu："当然，你可以现在美化图表。但我们才了解Xcelsius强大数据展现功能的一部分，在分析一些数据时总是希望图表还能显示与之相关的一些数据，数据钻取功能可以实现这个功能。"

（6）启动数据钻取功能

在【对象浏览器】窗格中双击【折线图】选项，显示【折线图1】对话框。单击【向下钻取】标签，显示【向下钻取】选项卡，如图5-16所示。



图5-16　【向下钻取】选项卡

选择【启用向下钻取】复选框，在【插入类型】下拉列表框中选择【值】选项，在【系列】列表框中选择【销售额】选项，在【目标】下拉列表框中选择一空单元格D4，存放钻取的数据。因为我们本例中只有销售额一个系列，所以无法钻取其他系列的数据，当数据系列较多时还可以选择钻取其他系列的数据。

设置后数据钻取到D4单元格，设置柱状图与折线图相同。数据钻取完成后它只能显示在Excel表格中，因此还需要展示。

（7）添加量表

在【部件】面板中选择【单值】选项，选择【量表】选项。双击

量表，显示【量表1】对话框，操作过程如图5-17所示。



图5.17　操作过程

将钻取的数据连接至【数据】下拉列表框，在【值范围】选项组中的【最小限制】文本框中输入“0”，在【最大限制】文本框中输入“2500”，即该量表能显示的数据范围。

单击【浏览】按钮显示图形，单击图形上的每个点可以在量表中显示相关数据。

水果商想知道的不仅仅是每月的销售数据，可能还想知道一年之中水果销售的平均值和最大值等数据，我们依然通过【量表】来显示。

重复以上步骤添加3个量表，分别显示最大值、最小值和平均值，并分别链接至Excel文件中的B7～B9单元格，如图5-18所示。



图5-18　链接单元格

在【量表2】对话框的【常规】选项卡中的【标题】下拉列表框中选择A8单元格，在【数据】下拉列表框中选择B8单元格，值范围设置为0和4000。

在【警报】选项卡中选择【启用警报】复选框，并选择【占最大值百分比】单选按钮。由于我们希望销售量越大越好，所以在【颜色顺序】选项组中选择【高值为好】单选按钮。

平均值与最小值的量表也按照此过程设置，设置完成后的结果如图5-19所示。

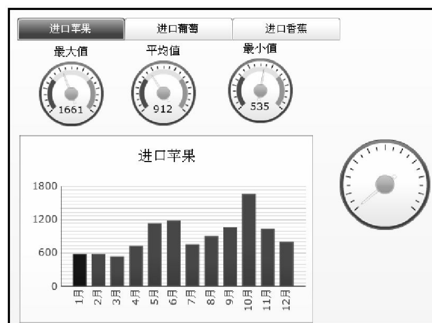


图5-19　设置完成后的结果

单击【浏览】按钮，动态图表的内容已经比较丰富，还可以添加Flash背景或图片进一步美化。

（8）插入Flash或图片

插入Flash文件或图片可以增加动态仪表盘的动感，在【部件】面板中选择【饰图和背景】选项。将【图像部件】选项拖动至画布中，并在【常规】选项卡中通过【导入】文本框导入所需图片，如图5-20所示。

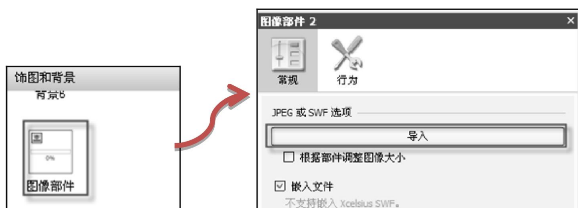


图5-20　导入图片

目前基本的部件都已添加完成，下面对整体进行美化。

（9）美化外观

我们可以根据对颜色的理解美化外观，Xcelsius也有强大的颜色组合方案供我们选择。

首先为部件添加背景，在【部件】面板中选择【饰图和背景】选项选择背景。将背景图片通过拖动覆盖所有部件。然后双击选择背景图片未覆盖画布的空白处，单击【使画布适应部件】按钮。单击工具栏中的按钮，使得背景与画布的大小一致。

将各个部件之间的布局和字体大小修订成需要展示的状态，然后选择【主题颜色】选项，显示的主题颜色如图5-21所示。

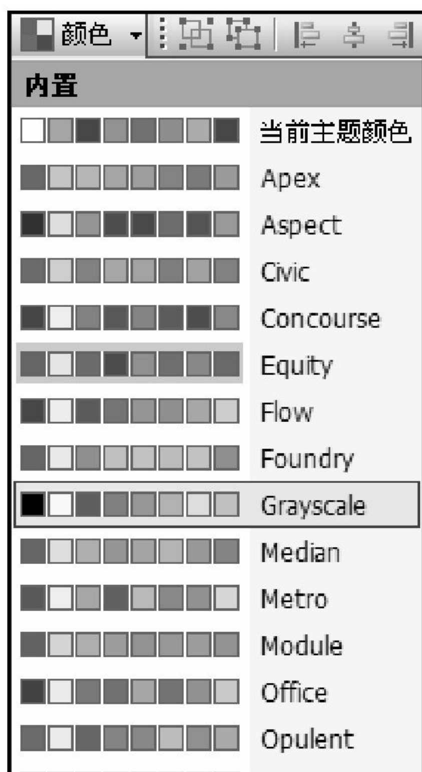


图5-21　主题颜色

根据展现的数据选择不同的颜色主题，并修订个别部件的设置。注意一旦选择颜色主题，之前保存的颜色设置不能恢复。

美化之后的效果如图5-22所示。



图5-22　美化之后的效果

此时可以导出这个漂亮的文档，在菜单中选择【文件】|【导出】|【Powerpoint】选项或PDF或其他格式。选择保存位置，在PDF中的展现效果如图5-23所示。

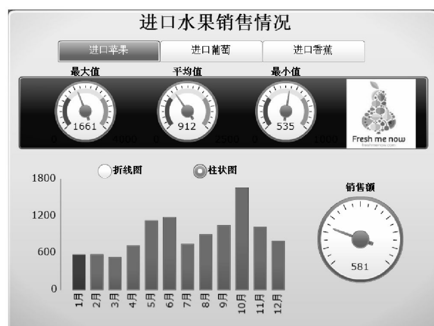


图5-23　在PDF中的展现效果

Miss　Ju："哇，这个报告真的是太漂亮了，如果拿它汇报工作或发邮件给领导，领导也会一目了然。"

Mr Shu："是的，一组简单的数据处理后可以产生无于伦比的精彩效果。但是美化的效果不仅需要自己满意，还要看是否能满足客户的要求。"

Miss　Ju：“在制作这个报告的过程中我发现有些部件还有很多功能尚未使用，它们应该也和前面一样的容易上手吧？”

Mr　Shu：“是的，通过上面的制作你应该知道Xcelsius其实很简单，它主要就是画布、部件和属性设置，属性设置功能比较强大。你还可以结合部件可以赋值的特点与Excel的功能结合，这样即可制作出非常实用又漂亮的数据报告。”

Miss Ju：“真的是没有做不到，只有想不到啊。”

Mr　Shu：“前面我们通过数据分析得出的结论在Excel中建好模型后即可与Xcelsius结合使用。还记得我们前面在讲解Crystal Ball软件时举的书店的例子吗？我们可以结合其数学模型在Xcelsius中展现。”

5.2　书店利润最大化

8月份书店需要确定订购明年日历的数量。单本日历的进价是7.5美元，售价是10美元。2月份之后所有未售的日历将会以2.5美元的价格退还给出版商。为了保证利润的最大化，如何来确定明年的日历数量？店主将以往的日历的销售数据进行了分布拟合，日历的销售量在100本～300本之间（均匀分布）。

在利用Crystal　Ball求最优解时我们需要整理出假设变量、决策变量与预测变量，在了解最优解之后再利用Xcelsius展现出来，为此需要根据Xcelsius软件的特点整理模型。

在求最优解时我们设置的假设变量为“需求量”，决策变量为“订货数量”，预测变量为“利润”，而“单位成本”、“单位售价”和“退货售价”这几个变量则是固定数据。

出于练习Xcelsius软件的需要，我们将“单位成本”、“单位售价”、“退货售价”、“需求量”和“订货数量”均设置为影响利润的变量，操作步骤如下。

（1）修订Excel模型

Crystal Ball中的数据模型如图5-24所示。

书店进书模型

数据

单位成本	$7.50
单位售价	$10.00
退货售价	$2.50

需求分布

最小需求	100
最大需求	300

模拟

订货数量	需求量	收入	成本	退款	利润
145	200	$1,450.00	$1,087.50	$0.00	$362.50

图5-24　Crystal Ball中的数据模型

在在C14单元格中输入公式“＝$B$5*MIN(A14,B14)”，即销售日历的收入；在D14单元格中输入公式“＝A14*$B$4”，即购买日历的成本值；在E14单元格中输入公式“＝MAX((A14-B14),0)*$B$6”，即退款的数量。

模型中的逻辑关系不变，只是将Crystal Ball中的变量设置去除。

（2）导入Excel数据

将“书店利润.xls”文件导入至Xcelsius中，启动Xcelsius软件。选择菜单中的【数据】|【导入】选项，导入“书店利润.xls”文件，导入后数据如图5-25所示。

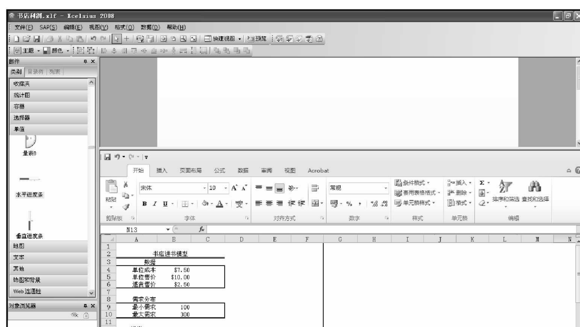


图5-25　导入的数据

（3）添加进度条

在【单值】部件中拖动【水平进度条】部件至画布中，如图5-26所示。

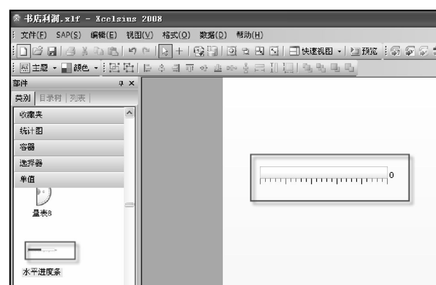


图5-26　添加水平进度条

双击【水平进度条】部件，显示【水平进度条1】对话框。在【常规】和【警报】选项卡中设置有关选项，如图5-27所示。



图5-27　设置有关选项

在【常规】选项卡的【标题】下拉列表框中引用F13单元格，在【数据】下拉列表框中引用F14单元格，将利润值的值范围调整至-100～500。在【警报】选项卡中选择【按值】单选按钮，并将红色设置为最小值0，黄色设置为0～250，绿色设置为250～最大值，在【颜色顺序】选项组中选择【高值为好】单选按钮。

设置完成后的水平进度条如图5-28所示。

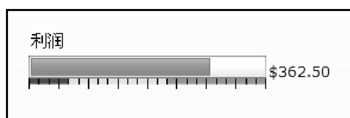


图5-28　设置完成后的水平进度条

此时Excel的初始值设置的利润值为362.5，水平进度条已显示并按照警报的颜色设置显示进度条为绿色。

为美化进度图外观，双击【水平进度条】部件，显示【水平进度条1】对话框。在【外观】选项卡中设置有关选项，如图5-29所示。



图5-29　设置有关选项

在【位置】下拉列表框中选择【中上】选项，字体加粗。在【数字格式】下拉列表框中选择【货币】选项，小数位设置为0，设置完成后的水平进度条如图5-30所示。



图5-30　设置完成后的水平进度条

（4）添加滑块

为了查看“单位成本”、“单位售价”、“退货售价”、“需求量”和“订货数量”5个数据对利润的影响，我们需要5个滑块。首先为“单位成本”添加一个滑块，在【单值】部件中选择【水平滑块】选项，添加过程如图5-31所示。

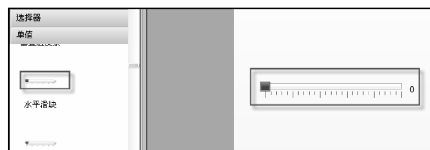


图5-31　添加滑块过程

双击【水平滑块】部件，显示【水平滑块1】对话框，如图5-32所示。



图5-32　【水平滑块1】对话框

水平滑块是根据滑块的滑动为Excel单元格赋值的一种行为。在【常规】选项卡中的【标题】下拉列表框中选择A4单元格，在【数据】下拉列表框中选择B4单元格。在【值范围】选项组中设置5～10（进货成本），即通过滑块向B5单元格赋5～10范围内的数值。

设置完成后的滑块如图5.33所示。

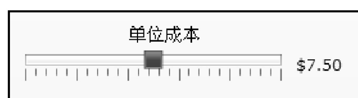


图5-33　设置完成后的滑块

由于还需要添加4个滑块，因此为了保持统一，设置“单位成本”滑块的外观，如图5-34所示。

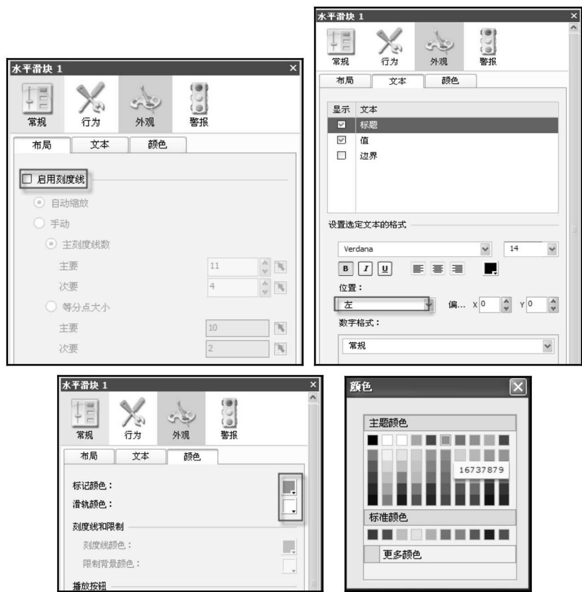


图5-34　设置水平滑块外观

在【常规】选项卡中清除【启用刻度线】复选框，在【位置】下拉列表框中选择【左】选项，在【值】下拉列表框中选择【右】选项，在【数字格式】下拉列表框中选择【右】选项“货币”选项，小数位为0。标记的颜色设置为色号为16737879的橙色，设置完成后的滑块如图5-35所示。

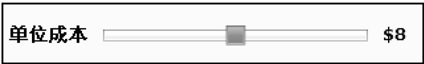


图5-35　设置完成后的滑块

在【对象浏览器】窗格中右击【水平滑块1】选项，选择快捷菜单中的【复制】选项。右击空白画布处，选择快捷菜单中的【粘贴】

选项4次，然后分别设置每个滑块。

设置的“单位售价”滑块如图5-36所示。



图5-36　设置的“单位售价”滑块

将【常规】选项卡中的【标题】下拉列表框中选择A5单元格，将【数据】下拉列表框中选择B5单元格，设置售价的范围为6～12。

设置“退货售价”滑块如图5-37所示。



图5-37　设置“退货售价”滑块

将【常规】选项卡中的【标题】下拉列表框链接至A6单元格，将【数据】下拉列表框链接至B6单元格，设置售价的范围为1～3。在【行为】选项卡中的【常用】选项组中设置为按0.5递增。

设置“订货数量”滑块如图5-38所示。



图5-38　设置“订货数量”滑块

在【常规】选项卡中的【标题】下拉列表框中选择A13单元格，在【数据】下拉列表框中选择A14单元格，设置售价的范围为80～300。

设置“需求量”滑块如图5-39所示。



图5-39　设置“需求量”滑块

在【常规】选项卡中的【标题】下拉列表框中选择B13单元格，在【数据】下拉列表框中选择B14单元格，设置售价的范围为80～320。

设置完成后的书店画布如图5-40所示。

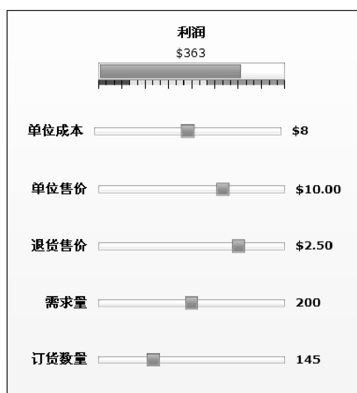


图5-40　设置完成后的书店画布

（5）添加背景及标题

为设置好的画布添加背景及标题，在【饰图和背景】面板中选择【背景4】部件作为背景并置于底层。

在画布中添加标题“书店销售利润”，如图5-41所示。

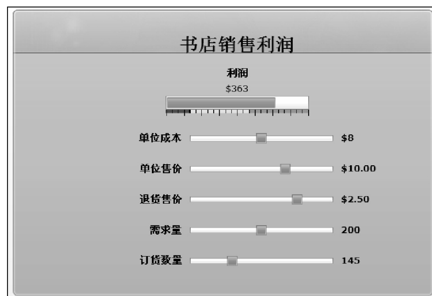


图5-41　设置的背景及标题

（6）美化外观

根据需求美化画布外观，颜色主题选择Concourse，美化后的画布如图5-42所示。

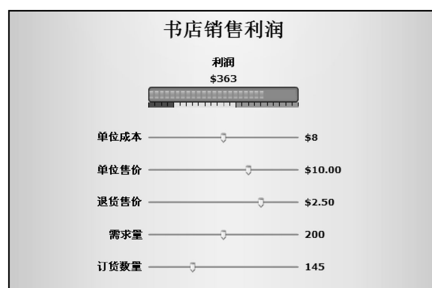


图5-42　美化后的画布

（7）发布报告

单击【浏览】按钮可以拖动滑块查看每个数据对利润的影响，然后可以根据需求导出为PDF或PPT文件。

Miss　Ju：“这样展现数据太直观了，可以看到每个数据对目标的影响。”

Mr　Shu：“这个案例运用了Excel和Xcelsius的一些功能，Xcelsius向Excel写入数据，Excel计算之后Xcelsius再展现数据。”

Miss Ju：“Xcelsius中的部件还有很多，与Excel结合好的话一定可以制作出更好且更炫的报告。”

Mr Shu：“开动灵活的大脑可以把报告做得更漂亮一些，学习了这么长时间的数据分析，你有哪些收获？”

Miss Shu：“Excel是基础啦，它本身也有很多数据分析功能，如规划求解、单变量求解和方案管理器等。它和Crystal Ball和Xcelsius两款软件的联系非常密切，学习好Excel，这两款软件也相对容易一些。在Excel的基础上我们还可以通过Minitab和JMP做一些相关性分析，如假设检验和回归分析等。有了模型后通过LINGO或Crystal Ball优化求解，有了优化结果通过Xcelsius展现分析成果，这样概括合适吗？”

Mr Shu：“进步神速嘛，你最喜欢哪一款软件？”

Miss Ju：“作为一名数据分析人员没有喜欢的，只有合适的，我认为这几款软件都是入门需要掌握的。但从个人喜好来讲，Xcelsius真的是太精彩了。Mr Shu，我们的课程是要结束了吗？Xcelsius的体验还没过瘾呢。”

Mr Shu：“差不多快要结束了，列举最后一个Xcelsius的案例之后结束我们的课程。”

Miss Ju：“案例的名称叫什么？”

Mr Shu：“你现在不是还是单身吗？不要参加什么《非诚勿扰》节目了，发挥你强大的数据分析能力帮自己选个中意的男友吧。”

Miss Ju：“你是说通过Xcelsius来选？嘻嘻，我最关心这个项目了，我们开始吧！”

5.3 非诚勿扰——最佳男友模型

Mr Shu：“你对男友有什么要求？如学历等。”

Miss　Ju：“基本的要求肯定是有的，不然没有共同语言，我们就按照世俗好男人的标准选择吧。”

Mr Shu：“那么按照基本条件和发展潜力两个维度来选择，基本条件包括身高、学历、体重和年龄等；发展潜力按照职业、月收入、房屋大小及车辆价位，你看行吗？”

Miss　Ju：“这也太庸俗了吧？还没开始就讨论房子啊什么之类的。”

Mr Shu：“这只是为了学习软件，不必太介意，现在我们先收集海选数据。”

（1）收集数据

收集100组备选男友的数据，如图5-43所示。

A	B	C	D	E	F	G	H
身高	体重	年龄	学历	职业	房产面积	收入	车辆级别
181	136	31	大专	高级经理	150	2390	高级车
165	123	23	硕士	普通经理	116	3642	低级车
167	159	24	本科	普通经理	104	14538	低级车
177	123	24	大专	职员	200	15942	中级车
168	167	30	本科	职员	168	5287	低级车
170	150	32	本科	普通经理	48	10655	普通车
185	160	31	硕士	金领	192	12507	低级车
179	164	23	硕士	职员	154	12657	中级车
166	169	28	博士	普通经理	110	8079	中级车
173	140	28	本科	金领	183	19032	高级车
176	126	26	博士	职员	56	13843	普通车
176	174	33	本科	金领	22	5959	普通车
173	171	28	本科	高级经理	49	14628	低级车
177	177	29	大专	高级经理	156	7984	低级车
185	159	33	博士	金领	25	12261	中级车
166	150	30	本科	职员	55	14821	高级车
174	144	26	硕士	职员	46	10924	普通车
181	161	33	博士	金领	197	3435	普通车
185	141	27	大专	职员	30	4890	中级车
174	160	32	大专	普通经理	59	16135	中级车
167	140	30	本科	高级经理	124	13894	普通车
181	173	27	博士	金领	81	16341	中级车
174	121	31	博士	普通经理	198	6192	低级车
170	126	23	大专	金领	127	7121	中级车
180	122	33	大专	职员	36	9581	普通车
178	170	24	博士	普通经理	92	16735	普通车
178	155	31	本科	金领	102	12745	中级车

图5-43　备选男友数据

Miss　Ju：“哇，这么多，但是为什么高级经理的收入却比不上一些职员？为什么有的职员的房产面积比经理还大？”

Mr　Shu：“这就是所谓的‘郎怕入错行’啊，行业与行业之间的收入差距非常巨大。至于房产，有的属于继承，即所谓的‘富二代’。”

Miss　Ju：“这样啊，我们怎么样来判断优质男友？仅看收入和房子肯定也不行啊。”

Mr Shu：“是的，这需要我们来拟定一个标准。例如，你比较看重哪个属性就将其权重调大一些。目前我们有8个属性，每个属性你都可以按照0～100评分。然后根据每个属性的权重进行加权，你看这样是否可以？”

Miss　Ju：“好像蛮有道理，那么我们现在设置每个属性的权重吧。”

（2）设置权重

设置8个属性的权重，如图5-44所示。

权重比例							
身高	体重	年龄	学历	职业	房产面积	收入	车辆级别
20%	10%	8%	6%	26%	10%	10%	10%

图5-44　设置8个属性的权重

Mr Shu：“看来你蛮看重职业和身高嘛。”

Miss　Ju：“还是职业最重要，事业是可持续的。身高不会变，其他都可以改变。”

Mr Shu：“有想法！设置各属性的权重后我们分配各属性的分值吧。”

（3）设置分值

Mr Shu："先从身高开始，你认为身高多少比较合适你？"

Miss　Ju："按照我自己的身高，男友的身高在173厘米～186厘米最好。"

身高设置如图5-45所示。

身高	分数
165	10
170	30
173	50
178	70
183	100
186	95
200	80

图5-45　身高设置

分别设置其他属性的分值，如图5-46所示。

设置分数							
身高	分数	体重	分数	年龄	分值	学历	分值
165	10	130	50	23	10	大专	40
170	30	140	70	25	40	本科	70
173	50	150	90	29	70	硕士	100
178	70	160	100	32	100	博士	60
183	100	170	80	35	40		
186	95	200	50	40	0		
200	80						
设置分数							
职业	分数	房产面积	分数	收入	分值	车辆级别	分值
职员	80	0	20	3000	10	低级车	80
普通经理	100	30	40	8000	50	普通车	100
高级经理	80	60	100	10000	80	中级车	80
金领	80	90	40	15000	100	高级车	80
		100	30	20000	100		

图5-46　设置的其他属性分值

身高165代表最低要求为165，170代表身高在165～170之间的分值为10；体重130代表体重的最低要求为130，140代表体重在130～240之间的分值为50，其他依次类推。

Mr Shu：“分配得很合理嘛，房子面积在60～90平米就是100分了，大房子反而分值不高。”

Miss Ju：“有点经济基础最好，后面的美好人生还是需要一起奋斗，我可不喜欢像寄生虫一样生活。”

Mr Shu：“分值和比例都分配完了，我们现在可以根据这些分值来计算分数了吧？”

Miss Ju：“是的，但是这里还有关键的一个步骤，即如何根据一个人的属性来计算其分值。例如，一个人的收入是10 000元。他在收入上的分值应该为80，这个过程如何来完成？”

Mr Shu：“在Excel中有很多函数都可以实现这个功能，在解决这个问题之前还需要提取原始数据导入到Xcelsius中。”

（4）设置数据区

在J24:Q26单元格区域设置原始数据存放区，为了更好地设置Excel函数，将原始数据表中的第1行数据复制到存放区，如图5-47所示。

原始数据存放区							
身高	体重	年龄	学历	职业	房产面积	收入	车辆级别
181	136	31	大专	高级经理	150	2390	高级车

图5-47　原始数据存放区

根据原始数据计算分值，在Excel中有较多的函数可以计算。本例选择Vlookup函数，在J30:Q32单元格区域中设置得分区，如图5-48所示。

数据得分区							
身高	体重	年龄	学历	职业	房产面积	收入	车辆级别

图5-48　数据得分区

在Excel中设置的函数如表5-1所示。

表**5-1**　设置的函数

单　元　格	公　　式	说　　明
J32	＝VLOOKUP(J26,J7:K13,2)	计算身高份的分值
K32	＝VLOOKUP(K26,L7:M13,2)	计算体重份的分值
L32	＝LOOKUP(L26,N7:O13)	计算年龄的分值
M32	＝VLOOKUP(M26,P7:Q11,2,FALSE)	计算学历的分值
N32	＝VLOOKUP(N26,J16:K20,2,FALSE)	计算职业的分值
O32	＝VLOOKUP(O26,L16:M20,2)	计算房产的分值
P32	＝VLOOKUP(P26,N16:O20,2)	计算收入的分值
Q32	＝VLOOKUP(Q26,P16:Q20,2,FALSE)	计算车辆的分值

设置函数后的各属性得分如图5-49所示。

数据得分							
身高	体重	年龄	学历	职业	房产面积	收入	车辆级别
70	50	70	40	80	30	#N/A	80

图5-49　数据得分

收入项的得分为错误值是因为我们设置的得分最低限为3 000，而该候选人收入只有2 390。即未达到最低限，所以会出现错误值。为了预防此类情况发生，设置数值低于最低限时判断分数为0，修正后的函数如表5-2所示。

表**5-2**　修正后的函数

单　元　格	公　　式	说　　明
J32	＝IF(J26＜＝165,0,VLOOKUP(J26,J7:K13,2))	计算身高份的分值
K32	＝IF(K26＜＝130,0,VLOOKUP(K26,L7:M13,2))	计算体重份的分值
L32	＝IF(L26＜＝23,0,VLOOKUP(L26,N7:O13,2))	计算年龄的分值

M32	＝VLOOKUP(M26,P7:Q11,2,FALSE)	计算学历的分值
N32	＝VLOOKUP(N26,J16:K20,2,FALSE)	计算职业分值
O32	＝VLOOKUP(O26,L16:M20,2)	计算房产的分值
P32	＝IF(P26＜＝3000,0,VLOOKUP(P26,N16:O20,2))	计算收入的分值
Q32	＝VLOOKUP(Q26,P16:Q20,2,FALSE)	计算车辆的分值

各项分数与权重相乘得到加权后的得分。为了分开“基本条件”和“发展潜力”两方面的分数，将8项属性分开统计，得分如图5-50所示。

数据得分区								
项目	身高	体重	年龄	学历	职业	房产面积	收入	车辆级别
分值	70	50	70	40	80	30	0	80
得分	14	5	6	2	21	3	0	8
大项得分	27				32			
总分	59							

图5-50　候选人得分

显然，第1名候选人的得分只有59分，尚未及格。

（5）设置画布

Mr Shu：“现在得到其中一名候选人的总分，其他人的得分我们也想知道。”

Miss　Ju：“有了这个得分，我们是不是应该设计如何展示数据？现在Excel模型已整理得差不多了吧？”

Mr Shu：“是的，下面我们设置Xcelsius画布。”

启动Xcelsius软件，将“精选男友.xls”文件导入其中。操作步骤与之前案例相同，导入的结果如图5-51所示。

图5-51　导入的数据

为了增加数据的交互效果，在【选择器】类别中选择【列表视图】选项，如图5-52所示。



图5-52　【列表视图】选项

将列表视图拖动至画布中，双击【列表视图】选项，设置属性如图5-53所示。



图5-53　设置属性

在【常规】选项卡中的【显示数据】下拉列表框中选择A2:H102单元格区域，在【数据插入】选项组中选择【行】选项。在【源数据】下拉列表框中选择A3:H102单元格区域，在【目标】下拉列表框中选择J26:Q26单元格区域。

即首先显示候选人的原数据，并将其选择至指定的原数数据存放区域的J26:Q26单元格区域。

在美化外观之前简单整理列表视图外观，如图5-54所示。



图5-54　整理列表视图外观

选择【自定义列宽】复选框，将每个列宽设置为50，如图5-55所示。



图5-55 设置列宽

将表头和标签的字体设置为11，表头字体加粗并居中，如图5-56所示。



图5-56 设置字体

选择【颜色】选项组中的【标签颜色】选项卡，设置标签颜色如图5-57所示。

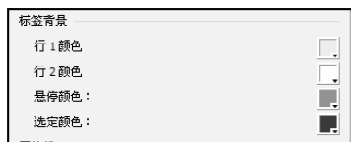


图5-57　设置标签颜色

将【悬停颜色】设置为色号为16737879的橙色，将【选定颜色】设置为第1个标准颜色，设置完成后的列表视图如图5-58所示。

身高	体重	年龄	学历	职业	房产面积	收入	车辆级别
181	136	31	大专	高级经理	150	2390	高级车
165	123	23	硕士	普通经理	116	3642	低级车
167	159	24	本科	普通经理	104	14538	低级车
177	123	24	大专	职员	200	15942	中级车
168	167	30	本科	职员	168	5287	低级车
170	150	32	本科	普通经理	48	10655	普通车
185	160	31	硕士	金领	192	12507	低级车
179	164	23	硕士	职员	154	12657	中级车

图5-58　设置完成后的列表视图

截至目前，我们已完成数据的提取工作，现在将提取的数据通过Excel计算后展现。

在【单值】面板中将【水平进度条】选项部件拖动至画布中添加水平进度条后复制。设置“身高”水平进度条，如图5-59所示。



图5-59　设置“身高”水平进度条

在【布局】选项卡中清除【启用刻度卡】复选框；在【文本】选项中将标题的字体位置设置为中上，大小设置为12，颜色设置为色号为16737879的橙色，将值的位置设置为右上；在【颜色】选项卡中设置标记颜色为色号为8089442，设置完成后将该进度条复制7个，如图5-60所示。

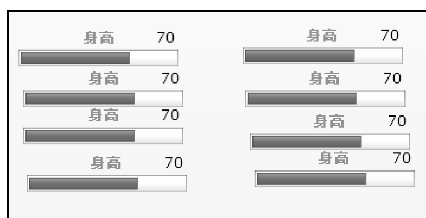


图5-60　设置外观后的进度条

将每个进度条分别链接至相关的数据和标题，并按照“基本条件”和“发展潜力”两个大类分开，如图5-61所示。

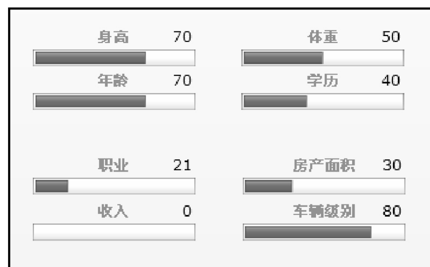


图5-61　进度条视图

将相同类别的属性使用连接线连接，在饰图和背景中选择【水平线】和【垂直线】选项连接。

将两个大类的得分情况通过“量表”来展现，在【单值】面板中选

择【量表】选项，添加量表后的画布如图5-62所示。

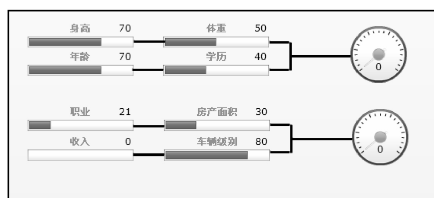


图5-62　添加量表后的画布

双击量表，显示【量表1】对话框，如图5-63所示。



图5-63　【量表1】对话框

在【常规】选项中的【标题】下拉列表框中输入“基本条件”，在【数据】下拉列表框中选择J34:M34单元格区域，即“基本条件”的得分情况。将【值范围】选项组的最小值设置为0，最大值设置为K4单元格。在Excel模型中K4单元格设置为“＝SUM(J3:M3)*100”，即“基本条件”的总比例。在N4单元格中设置公式“＝SUM(N3:Q3)*100”，即“发展潜力”总比例。在【警报】选项卡中设置有关选项，如图5-64所示。

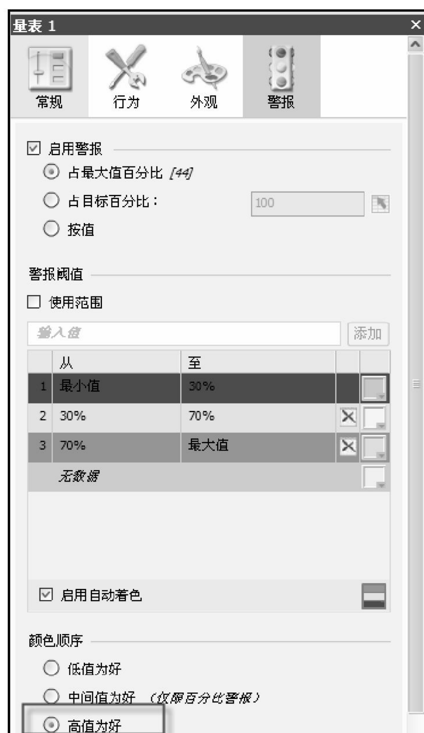


图5-64　设置有关选项

选择【启动警报】复选框和【占最大值百分比】单选按钮，在【颜色顺序】选项组中选择【高值为好】单选按钮。

双击另一个量表，设置“发展潜力”量表，完成后的量表如图5-65所示。

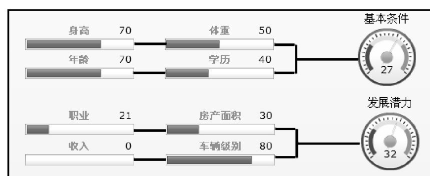


图5-65　设置完成后的量表

添加“基本条件”与“发展潜力”后还需要设置综合总得分，设置过程与之前一致，设置完成后的综合总得分，如图5-66所示。

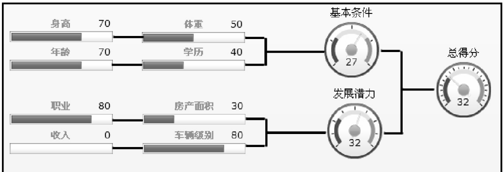


图5-66　设置完成后的综合总得分

（6）浏览

截止到目前，已设置完成该模型的雏形，单击【预览】按钮可以预览文件。此时使用鼠标选择数据表中的数据可以实现实时展现各项数据得分，如图5-67所示。

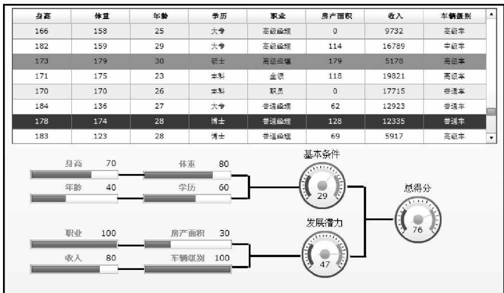


图5-67　预览视图

Miss　Ju：“哇，这样看得好清楚。好有成就感。这样选男友既理性，也有趣。”

Mr Shu：“是的，但是好像有点问题。按照我们之前的标准，很多优秀男的得分是不是有点低？查看了一下最高分好像也只有76分。”

Miss　Ju："那怎么办？重新制定标准吗？那岂不是每次都要修改Excel模型？"

Mr　Shu："当然不需要，上一节的案例不就是动态修正参数吗？动态修正参数可以通过设置控件实现。"

（7）设置参数控件

在【单值】面板中选择【刻度盘7】部件并添加至画布中，双击该控件，设置有关选项，如图5-68所示。

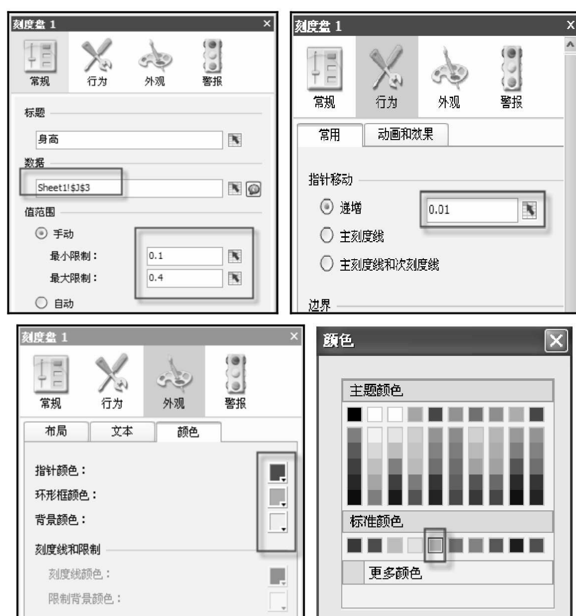


图5-68　设置有关选项

在【标题】下拉列表框中输入"身高"，在【数据】下拉列表框中选择J3单元格。将【值　范围】设置为0.1～0.4，指针和环形颜色分别选择标准颜色卡中的第2个和第5个。设置完成后复制7个刻度盘，分别对应"体重"和"年龄"等属性，添加的刻度盘如图5-69所示。

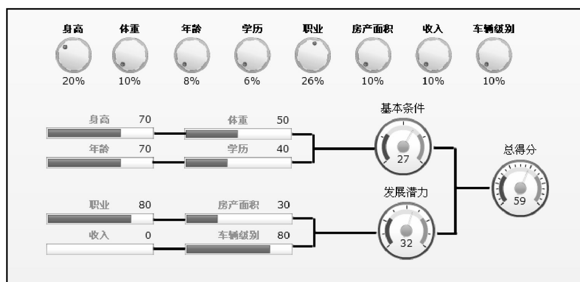


图5-69　添加的刻度盘

现在预览，不需要修改参数就可以直接查看结果。如果修改权重，可以直接在刻度盘上通过拖动修改，修改后的数据也可以直观地查看。

Miss　Ju：“每个刻度盘设置的值的范围不一样啊，有的是0.1～0.4；有的是0.08～0.2等，万一分配的权重加起来超过100%怎么办？”

Mr　Shu：“观察得很仔细嘛，的确可能超过100%。为了防止这种情况发生，可以在Excel中设置犯错条件或在画布中添加预警。”

Miss　Ju：“在Excel中设置犯错条件指必须要求8项的和为100%，在画布中添加预警指通过报警来提示吗？”

Mr　Shu：“是的，选择其中一项在Excel中将数据修正为公式。”

在Q3单元格中输入公式“＝1-SUM(J3:P3)”，即将“车辆等级”这个属性作为观察项确保8项属性的权重和为100%。双击“车辆等级”刻度盘并启动警报，设置有关选项如图5-70所示。



图5-70　设置有关选项

在【常规】选项卡中设置值的最小范围为-0.5，在【警报】选项卡中选择【启用警报】和【按值】单选按钮预警，清除【启动自动着色】复选框。从最小值到0设置为红色，0到最大值为绿色，设置完成后该刻度盘的数大于0时属性的权限之和为1。

（8）美化外观

Mr Shu：“是不是迫不及待想看看效果？截止到现在，该模型的功能均已设计完成，现在剩下的工作就是美化外观了。”

Miss Ju：“是的，我想看看哪些符合自己的标准。”

Mr Shu：“美化外观是一个个性化的过程，给出的效果也只是一个参考。”

美化外观后的效果如图5-71所示。



海选男友模型

身高	体重	年龄	学历	职业	房产面积	收入	车辆级别
181	136	31	大专	高级经理	150	2390	高级车
165	123	23	硕士	普通经理	116	3642	低级车
167	159	24	本科	普通经理	104	14538	低级车
177	123	24	大专	职员	200	15942	中级车
168	167	30	本科	职员	168	5287	低级车
170	150	32	本科	普通经理	48	10655	普通车
185	160	31	硕士	金领	192	12507	低级车
179	164	23	硕士	职员	154	12657	中级车

图5-71　美化外观后的效果

现在预览，可以动态选择每个人的得分，还可以动态分配每个参数的权重；另外，单击列表中的表头可以排序数据，如降序排列身高。单击表头“身高”，结果如图5-72所示。

身高	体重	年龄	学历	职业	房产面积	收入	车辆级别
185	160	31	硕士	金领	192	12507	低级车
185	167	26	博士	高级经理	0	6573	中级车
185	141	27	大专	职员	30	4890	中级车
185	159	33	博士	金领	25	12261	中级车

图5-72　降序排列身高

Mr　Shu：“哇哦，是不是可以非常清晰地筛选出你的白马王子？”

Miss　Ju：“嘻嘻，筛选白马王子只是一个操作。关键是可以学习到水晶易表的强大功能，我们现在是不是可以导出为PPT或PDF格式？”

Mr Shu：“可以，讨论到现在我们的数据分析入门课程就算结束了，下面做一个简短回顾。”

第1章从不靠谱的世界杯预测开始接触概率分布的概念，通过部

分案例了解了概率分布在定量风险分析和一些常用的概率分布，以及分布之间的关联。并且介绍了Minitab在分布模拟的一些功能，以及Crystal Ball在预测和定量风险分析的应用。

第2章从Excel本身的数据分析功能出发，通过理财案例学习这些功能，如规划求解和单变量求解等。随后学习了线性规划的基本原理，以及Crystal Ball和LINGO在优化求解方面的强大功能。

第3章我们重点了解了JMP软件在图形方面的强大功能，重点介绍热力图和SPC图，以及SPC图之间的关联。

第4章讲解了一些比较枯燥的数据分析的方法，如假设检验、相关与回归和聚类等。尽管枯燥，我们也尽量列举一些简单有趣的案例。以尽可能易懂的方法说明这些分析方法，主要的工具是Minitab。JMP在这方面的功能与Minitab类似，有些甚至更加强大。

第5章说明数据展现，主要讨论Xcelsius强大的数据展现功能，Xcelsius需与Excel结合实现这些功能。只要思路正确，Xcelsius可以做出一些效果让人叹为观止的报告。

综合来讲，Excel是数据分析入门的基础。而Minitab、JMP、Xcelsius和Crystal Ball是业界的精英软件，它们以业务为核心，灵活多变地运用它们，一定会有所获。

彩插



图1-17 【B9单元格】



图1-27 计算结果

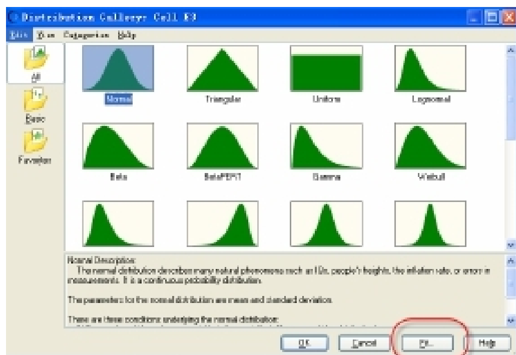


图1-52 【Fit】选项

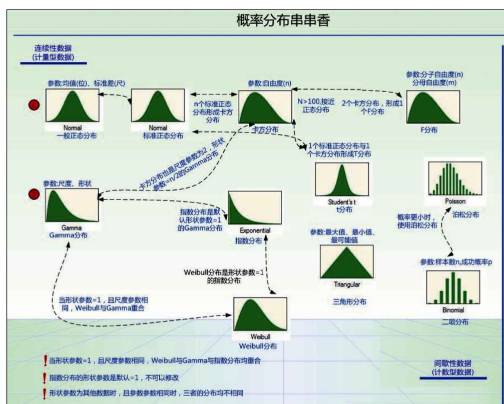


图1-93 常用的几种分布

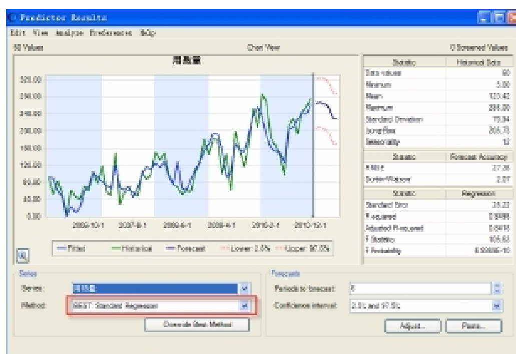


图1-106 回归预测分析结果



Objectives	Value
Minimize the Mean of 总成本	225,300.00

Requirements	Value

Constraints	Left Side		Right Side
CB模型!C17 <= 400	280	<=	400
CB模型!H22 <= 500	470	<=	500
CB模型!H23 <= 600	500	<=	600
CB模型!I21 <= 300	230	<=	300
CB模型!I22 <= 300	280	<=	300
CB模型!I23 <= 300	250	<=	300
CB模型!G17 >= 400	400	>=	400
CB模型!G18 >= 200	200	>=	200

Decision Variables	Value
六月322A量	5,000.00
六月802B量	0.00
四月322A量	400.00
四月802B计划量	3,000.00
五月322A量	3,500.00
五月802B量	1,500.00

图2-87　优化计算结果



图2-103　敏感度图

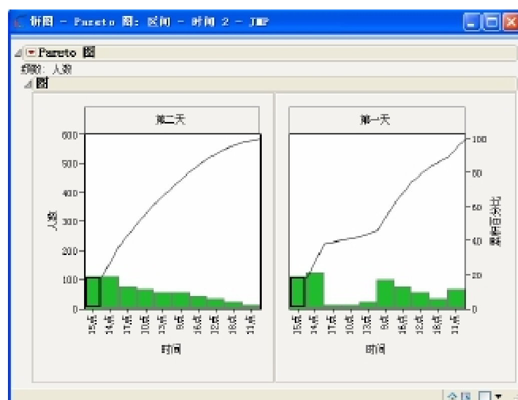


图3-17　增加X 分组后的Pareto图

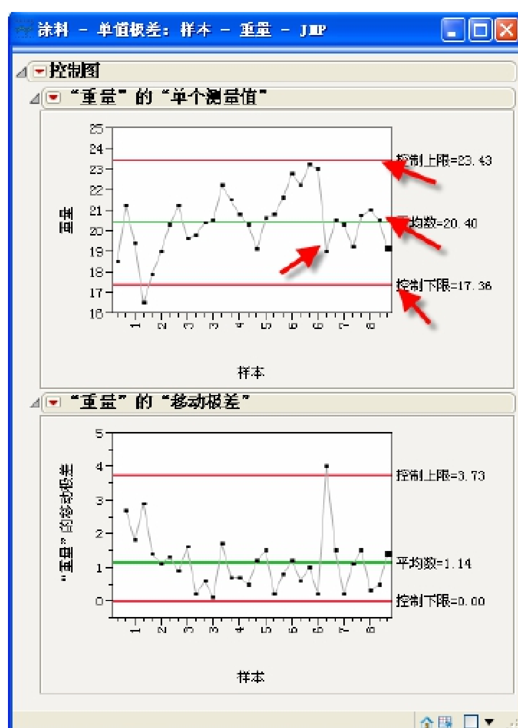


图3-41　生成的单值极差图

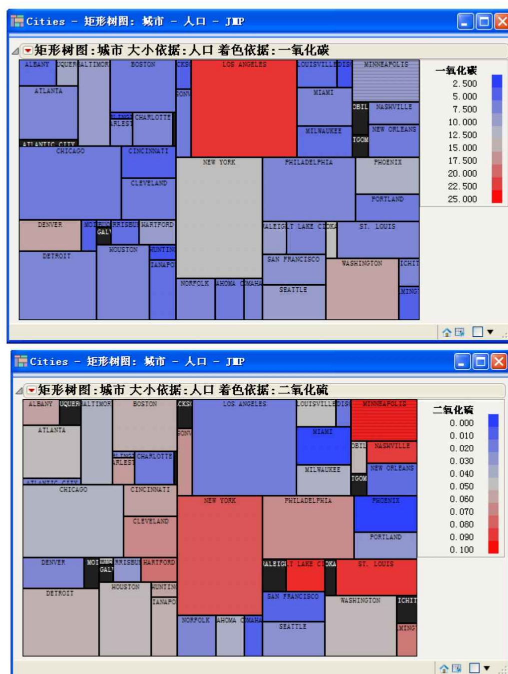


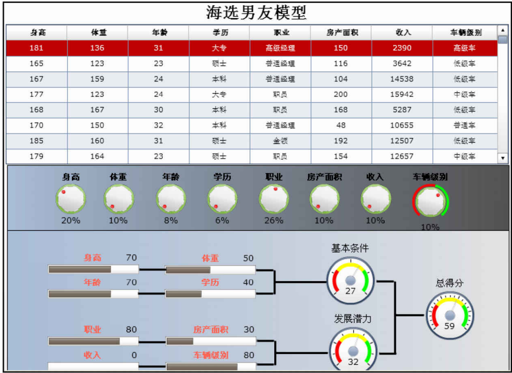
图3-71　着色之后矩形树图



图3-71　着色之后矩形树图（续）



图5-23　在PDF中的展现效果



海选男友模型

身高	体重	年龄	学历	职业	房产面积	收入	车辆级别
181	136	31	大专	高级经理	150	2390	高级车
165	123	23	硕士	普通经理	116	3642	低级车
167	159	24	本科	普通经理	104	14538	低级车
177	123	24	大专	职员	200	15942	中级车
168	167	30	本科	职员	168	5287	低级车
170	150	32	本科	普通经理	48	10655	普通车
185	160	31	硕士	金领	192	12507	低级车
179	164	23	硕士	职员	154	12657	中级车

图5-71　美化外观后的效果